COUNTWAY LIBRARY
HC 2461 %

ADVANCES IN MEDICAL PHYSICS

John S. Laughlin, *Editor*

Edward W. Webster, *Associate Editor*

I.C.M.P.

Symposium Papers of the
Second International Conference on Medical Physics

BOSTON
MEDICAL LIBRARY

IN THE
Francis A. Countway
Library of Medicine
BOSTON

Gift of
New England Journal of Medicine

A

ADVANCES
IN MEDICAL PHYSICS

ADVANCES IN MEDICAL PHYSICS

Symposium Papers of the
Second International Conference on Medical Physics

Boston, Massachusetts, U.S.A.
August 11-15, 1969

John S. Laughlin, *Editor*
Sloan-Kettering Institute and Memorial Hospital, New York, N.Y.

Edward W. Webster, *Associate Editor*
Massachusetts General Hospital, Boston, Mass.

I.C.M.P.

Published by
The Second International Conference on Medical Physics, Inc.
Boston, Massachusetts, U.S.A.

1971

FRANCIS A. COUNTWAY
LIBRARY OF MEDICINE

1971 © Second International Conference on Medical Physics, Inc.
Boston, Massachusetts, U.S.A.

Library of Congress Catalog Card Number: 78-167751

QT
34
I647
1969a
copy 1

Acknowledgments

The Second International Conference on Medical Physics was supported by the following agencies in the United States:

U.S. Department of Health, Education, and Welfare, Environmental Control Administration, Bureau of Radiological Health, Grant Number #EC 00130-01.

U.S. Atomic Energy Commission, Division of Technical Information, Contract Number AT(49-7) 3033.

The American Cancer Society, Grant Number R45.

Printed and bound by McGregor and Werner, Inc.
6411 Chillum Place, N.W., Washington, D.C. 20012 U.S.A.

Further copies are obtainable from:
The Second International Conference on Medical Physics, Inc.
Secretary, Dr. E. W. Webster, Department of Radiology
Massachusetts General Hospital
Boston, Massachusetts, U.S.A. 02114

THE SECOND INTERNATIONAL CONFERENCE ON MEDICAL PHYSICS

Boston, Massachusetts, U.S.A.

August 11-15, 1969

under the auspices of

THE INTERNATIONAL ORGANIZATION
FOR MEDICAL PHYSICS

and

THE U.S. NATIONAL COMMITTEE
ON MEDICAL PHYSICS

Honorary President:	Edith H. Quimby
President:	Lauriston S. Taylor
Secretary-General:	Edward W. Webster

Scientific Program Committee:

Chairman:	John S. Laughlin
	Gordon L. Brownell
	R. Stuart Mackay
	Herbert M. Parker
	Harald H. Rossi
	Michel M. Ter-Pogossian

THE INTERNATIONAL ORGANIZATION FOR MEDICAL PHYSICS

1965-1969

President	*Vice-President*	*Sec.-General*
W. V. Mayneord U.K.	J. S. Laughlin U.S.A.	B. Waldeskog Sweden

1969-1972

President	*Vice-President*	*Sec.-General*
J. S. Laughlin U.S.A.	R. I. Magnusson Sweden	J. R. Cameron U.S.A.

THE U.S. NATIONAL COMMITTEE ON MEDICAL PHYSICS

W. K. Sinclair, *Chairman*
G. L. Brownell
J. R. Cameron
R. O. Gorson
J. Hale
W. J. MacIntyre
M. Ter-Pogossian
E. W. Webster
M.M.D. Williams
P. Wootton

The following scientific societies are represented on the above Committee and are the U.S. sponsors of the Conference:

THE AMERICAN ASSOCIATION OF PHYSICISTS IN MEDICINE
THE HEALTH PHYSICS SOCIETY
THE RADIATION RESEARCH SOCIETY
THE SOCIETY OF NUCLEAR MEDICINE

FOREWORD

PHYSICS IN MEDICINE

The question is often asked, how physics relates to medicine and where does the physicist fit into the carrying out of medical practice and procedures? The practice of medicine is not really very scientific in the strict sense. Yet there are countless scientific pieces of information that provide a part of our medical knowledge and which could contribute essentially to the better conduct and understanding of the medical problem; many of these pieces are what might be described as coming in the category of physics—or, engineering. It is interesting and significant that in the biophysics discipline, most of the early scientists entered by way of the physics door—having first learned the disciplines of a hard science and then sought ways to apply these techniques to biological or biomedical problems.

No better example can be found of how the exact science of physics can contribute to the solution of biomedical problems than the series of seven symposia held during the 2nd International Conference on Medical Physics. Twenty-five speakers, invited from around the world, demonstrated the application of physical and engineering techniques and principles to the accomplishment of tasks important to medical diagnoses or treatments, and it is those papers that are selected and presented in this volume.

But even though physics may be regarded as an exact science, its application in medicine seldom leads to unequivocal and firm answers. This is in part because biology in medicine is not yet the exact science—and the cause-effect relationships that are the backbone of physics rarely exist in medicine. Thus many of the better known base points have to be based upon statistical evidence and this may well imply a large number of experiments to establish a single point on a curve. There is also the fact that biological experiments may directly or indirectly affect large numbers of people and hence unusual caution must be exercised in the promotion of research ideas when there may be even a small chance of having an adverse effect on the final experimental medium—man.

It is because of these two facts that value judgements play an increasingly important role in medical physics and its offshoots—like radiological physics.

Medical advances—in some of which physics has played a part—have been so great that within our lifetime, man's life expectancy has been extended by a decade. This has been attended by many new problems of medical, social and economic natures. Coupled with the present pace in the world population growth, these problems will reach some point where the effort needed to cope with them will be greater than our foreseeable ability to cope with them. What the medical physicists can do to aid in their solution I do not know, but I do know that he cannot hide himself from the problem, and along with the medical profession this should be very much on his mind.

Lauriston S. Taylor
President, Second International Conference on Medical Physics
Washington, D.C., 1971

PREFACE

The papers gathered together in this volume are those invited for presentation in the symposia held during the Second International Conference on Medical Physics. All of these papers were selected by the Program Committee for the Conference, the members of which worked diligently over a long period of time to organize the scientific content of the Conference. This volume would not have been possible without the joint activities of the authors of these papers and of the Program Committee members.

The scientific objectives of the Conference were two-fold: to illustrate the scope of medical physics; and to demonstrate the quality which had been achieved in each of the many applications of physics in medicine. One of the earliest and most major applications of physics in medicine resulted from the discovery of x-rays by Roentgen in a physics laboratory. This discovery has been applied to both therapeutic and diagnostic medicine as well as to research. As indicated by the distribution of papers at this Conference, the preponderance of effort by medical physicists is still related to the diagnostic and therapeutic uses of ionizing radiation. However, many papers in the Conference were related to other applications of physics and this activity was reflected in its selection of symposium papers by the Program Committee. The new area of nuclear medicine was specifically recognized through a symposium and is a further instance of an important medical field, primarily diagnostic, which has had its origin in a number of physics laboratories and in which physicists are continuing to play a primary role. The symposium on Measurement of Human Physiological Responses illustrates important medical applications other than the diagnostic and therapeutic uses of x-rays. The papers presented in the two New Frontiers symposia on holography, cardiac assistance, scanning electron microscopy, and the symposium on Ultrasound are further illustrations of medical physics unrelated to ionizing radiation. The last symposium on Information Processing describes a particularly significant area in which physicists have employed mathematical techniques and computer technology in the interests of medicine.

The Program Committee also wishes to acknowledge the contributions to the Second International Conference on Medical Physics by two other bodies. The U.S. National Committee for Medical Physics initiated the planning of the Conference. The International Organization for Medical Physics (IOMP) was the international sponsor for the Conference and plays a continuing role in promoting international cooperation between medical physics organizations. The First International Conference on Medical Physics was convened by the IOMP in September, 1965, in Harrogate, England. With the scope and quality of work evidenced here, we look forward to succeeding International Medical Physics Conferences sponsored by the International Organization for Medical Physics.

John S. Laughlin
Chairman, Program Committee

LIST OF CONTRIBUTORS

W. G. Austen, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts
M. Berman, National Institutes of Health, Bethesda, Maryland
G. F. Bond, U.S. Naval Deep Submergence System, Chevy Chase, Maryland
M. J. Buckley, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts
J. R. Cameron, University of Wisconsin, Madison, Wisconsin
J. H. Carrington, University of Southern California, Los Angeles, California
J. S. Clifton, University College Hospital, London, England
W. J. Collis, Northwestern University, Evanston, Illinois
T. D. Cradduck, Foothills Hospital and University of Calgary, Alberta, Canada
J. H. Cundiff, The University of Texas M.D. Anderson Hospital and Tumor Institute, Houston, Texas
E. W. Emery, Royal Postgraduate Medical School, London, England
E. Feleppa, Riverside Research Institute, New York, New York
R. Golden, The University of Texas M.D. Anderson Hospital and Tumor Institute, Houston, Texas
R. A. Greenes, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts
T. L. Hayes, Donner Laboratory, University of California, Berkeley, California
J. G. Holt, Memorial Hospital, New York, New York
H. E. Johns, Ontario Cancer Institute and University of Toronto, Ontario, Canada
R. T. Jones, Avco Everett Research Laboratory, Everett, Massachusetts
J. M. Jurist, University of Wisconsin, Madison, Wisconsin
A. R. Kantrowitz, Avco Everett Research Laboratory, Everett, Massachusetts
A. M. Kellerer, Columbia University, New York, and University of Munich, West Germany
C. Kellershohn, Commissariat à l'Energie Atomique, Orsay, France
J. D. Laird, Avco Everett Research Laboratory, Everett, Massachusetts
R. S. Mackay, Boston University, Boston, Massachusetts
P. N. Madras, Avco Everett Research Laboratory, Everett, Massachusetts
R. Martin, University of Southern California, Los Angeles, California
B. H. Mayall, University of Pennsylvania, Philadelphia, Pennsylvania
R. B. Mazess, University of Wisconsin, Madison, Wisconsin
M. L. Mendelsohn, University of Pennsylvania, Philadelphia, Pennsylvania
J. M. Morgan, The Radiation Center, Fort Worth, Texas
N. Palley, University of Southern California, Los Angeles, California
D. Parker, University College Hospital, London, England
B. H. Perry, University of Pennsylvania, Philadelphia, Pennsylvania
S. M. Pizer, Massachusetts General Hospital, Boston, Massachusetts, and University of North Carolina, Chapel Hill, North Carolina

H. R. Schreiner, Ocean Systems, Inc., Tarrytown, New York

R. J. Shalek, The University of Texas M.D. Anderson Hospital and Tumor Institute, Houston, Texas

H. Shubin, University of Southern California, Los Angeles, California

C. N. Smyth, The City University and University College Hospital, London, England

M. Stovall, The University of Texas M.D. Anderson Hospital and Tumor Institute, Houston, Texas

M. M. Ter-Pogossian, Washington University Medical School, St. Louis, Missouri

B. J. Thompson, University of Rochester, Rochester, New York

C. A. Tobias, University of California, Berkeley, California

A.C.C. Tseung, City University, London, England

M. H. Weil, University of Southern California, Los Angeles, California

H. S. Wolff, National Institute for Medical Research, Holly Hill, London, England

TABLE OF CONTENTS

Title Page and Acknowledgments .... iii, iv
Organization .... v, vi
Foreword vii
Preface viii
List of Contributors ix

Symposium A, Physical Aspects of Radiobiology

Chairman, Harald H. Rossi

The Time Factor in Radiobiology, E. W. Emery 1
Statistical Aspects of Radiation Action, A. M. Kellerer 13
Physical Energy Transfer and Biological Effect, C. A. Tobias 28

Symposium B, Measurement of Human Physiological Responses

Chairman, Herbert M. Parker

Galvanic Cell Transducers for the In-Vivo Measurement of Oxygen Tension, J. S. Clifton 51
Measurement of Human Physiological Responses, H. S. Wolff 65
Some Physical Methods of Skeletal Evaluation, J. R. Cameron 79
Automated Techniques for the Bedside of the Critically Ill, M. H. Weil 91

Symposium C, New Frontiers I

Chairman, Edward W. Webster

Holography and Its Application to Medical Physics, E. Feleppa 99
Cardiac Assistance by the Intra-Aortic Balloon, A. R. Kantrowitz 112
The Scanning Electron Microscope: Multi-Radiation, Non-focused Imaging, T. L. Hayes 124

Symposium D, Physical Aspects of Nuclear Medicine

Chairman, John S. Laughlin

New Formulation of Earlier Results in the Kinetic Theory of Tracers. Application to Various Problems in Metabolism, C. Kellershohn .. 141
Modulation Transfer Function and Its Application to Radioisotope Scanning, T. D. Cradduck 164
An Image Intensifier Isotope Camera Capable of Pulse Height Discrimination, M. M. Ter-Pogossian 179

Symposium E, Ultrasound

Chairman, R. Stuart Mackay

Ultrasonic Imaging with Image Cameras, Holography and Colour-Phase Representation, C. N. Smyth 185
Ultrasonic Imaging and Decompression Sickness Studies, R. S. Mackay 204

Symposium F, New Frontiers, II

Chairman, Michel M. Ter-Pogossian

Photochemistry of Excited States of DNA Components, H. E. Johns . . 217
Physical Limits of Manned Underwater Activity, H. R. Schreiner 232
Man's Role in Inner Space, G. F. Bond 255

Presentation on Professional Aspects of Medical Physics and Interdisciplinary Relationships

Cooperative Programs and the Status of Medical Physics, R. J. Shalek 273

Symposium G, Information Processing

Chairman, Gordon L. Brownell

Compartmental Modeling, M. Berman 279
Resolution Recovery in Radioisotope Scans, S. M. Pizer 297
Application of Computers to Treatment Plan Radiation Dose Computation, J. G. Holt 305
Optical Image Processing, B. J. Thompson 315
Generalized Grayness Profiles As Applied to Edge Detection and the Organization of Chromosome Images, M. L. Mendelsohn 327
A Model of a Computer-Based System for Medical Record-Keeping by Physicians, R. A. Greenes 342

THE TIME FACTOR IN RADIOBIOLOGY

E. W. Emery

It is usual to specify an irradiation in the first place by the absorbed dose, the energy imparted per unit mass of material. A complete specification would involve stating how many particles, of what track density, had crossed unit area in the material irradiated; how the tracks were distributed in time, and possibly also in direction. To determine all these parameters in practice would be an enormous labour, and in our present state of knowledge we should be unable to make use of more than a little of the mass of information that it would yield. We therefore relate radiation effects in the first instance to absorbed dose, and regard the differences between the various irradiations that would yield the same value for absorbed dose as modifying factors. This practice tends to give energy absorption a somewhat exaggerated importance; for many purposes it might be more informative to know the total track length per unit volume, and to regard the other parameters of the irradiation as affecting the response per unit track length. However, absorbed dose is the easier quantity to measure, and, although the response for a given dose may vary by a factor often greater than two and sometimes greater than ten, it is still probably a more slowly varying function of the other parameters than would be the response for a given value of any other quantity we could specify. Nevertheless, the well known statement that ionising radiation deposits energy in the material, which brings about the radiation changes, conduces to some dangerous short cuts in thinking. It should not be forgotten that the first effect of the radiation ". . . is to produce excitations and ionizations in molecules of the medium, and . . ." that chemical changes follow from these, leading ultimately to the observed effects. However, the yield of each of the possible excitations is, subject to statistical variations, proportional to energy absorbed almost irrespective of the type and velocity of the charged particle. Response per unit energy absorbed is therefore, almost fortuitously, a good concept to work with.

Of the factors influencing the response per unit absorbed dose the best known and most thoroughly investigated is the density of energy transfer along the track, the LET. It is well known that a given absorbed dose from high LET radiation often has the biological effect of several times that dose absorbed from low LET radiation. RBE tends to increase with LET, at least up to 100 to 200 kev/μm by the conventional definition of LET. Absorbed energy is therefore more effective if it is concentrated. One is naturally led to consider what is the volume within which the concentration is effective; that is, what is the volume of the "target". The average energy deposition when a particle of LET L keV per μm crosses a sphere of radius $\mathbf{a}\,\mu$m is $(4/3)\mathbf{a}L$ keV. Averaged over the sphere, this is the same energy deposition that would be expected from an absorbed dose of $5.1\ L/\mathbf{a}^2$ rads. Table 1 shows the equivalent dose for spheres of various sizes crossed by a track of 50 keV/μm. If this dose were delivered as X or γ radiation, each target area would receive its energy deposition from about 100 tracks, and the statistical variations would therefore be quite small. It might be expected therefore that if a dose of X-rays could be given extremely quickly, its RBE would increase towards that of radiation of much higher LET. By noting the dose at which the RBE became equivalent to that of radiation of, say, 50 keV/μm, one could interpolate in Table 1 to find the target volume. Then by decreasing the dose rate at which the X-rays were delivered until the RBE reverted to the normal X-ray value, one could learn about the time scale of the interactions between the immediate products of radiation that underlie the enhanced RBE. These attractive possibilities must have engaged the thoughts of numerous workers, but unfortunately the simple experiments outlined above are not feasible. As Table 1 shows, if the target volume is much smaller than a cell, the dose involved is very large. For instance, the work of Barendsen et al.(1) on the killing of kidney cells in tissue culture indicates a target size of the order of 0.01 μm, suggesting that a pulsed X-ray dose of some mega-rads would be necessary to reveal an enhanced RBE. At such doses, the number of surviving cells would be negligibly small*. There are of course also the

Table 1. Energy Deposition in Unit Density Spheres

Radius of sphere (μm) (density = 1 gm cm^{-3})	*Absorbed dose to give same energy * deposition as one traversal by a particle of LET 50 keV/μm (rads)*
10	2.5
1	250
0.1	25,000
0.01	2,500,000

* A more appropriate quantity is the parameter Z as defined by Rossi (17). Values of Z in ergs/g are 100 times the figures appearing in this column.

---

* The paradox that the dose required to place an average of one high LET track in each target is many times the dose required to produce 37% survival, can be resolved if it is assumed that each cell contains many targets. The targets could, of course, be in some organized form such as a continuous lamina.

technical problems of delivering the appropriate doses. The most purely physical way in which time might enter radiobiology is, in principle, this interchangeability of dose rate and LET. But technical difficulties are such that it does not appear to have been demonstrated in practice. It is tempting to speculate how the difficulties might be overcome, but in the absence of experimental results, such speculation is better left as a matter for private recreation.

Effects of Very High Dose Rate

At the present time experiments in which low LET radiation has been delivered at very high dose rate show that, far from increasing the effectiveness of the radiation the high dose rate leads to a diminution of effectiveness as shown in Fig. 1, which gives results obtained by Epp et al. (2). The effect was first observed with bacteria by Dewey and Boag (3) and has been observed by other workers (4, 5, 6, 7, 8).

PULSED IRRADIATION OF CELLS IN OXYGEN

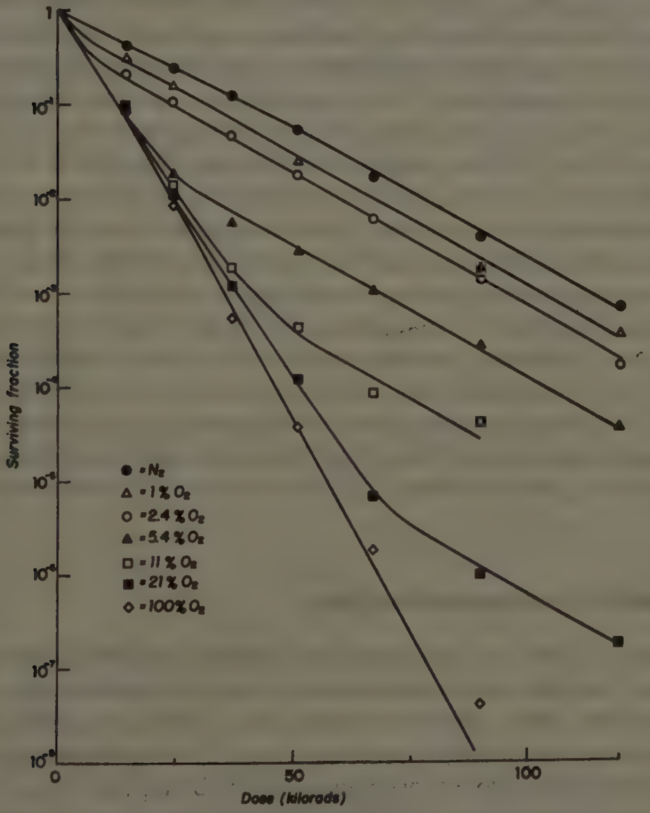


Fig. 1. Survival of bacteria (E *coli* B/r) irradiated with single high intensity pulses of electrons of approximately 30 ns duration (Epp et al. (2)).

The effect is the result of chemical rather than physical mechanism. As the curve for pulsed radiation breaks away from that obtained with normal dose rates it assumes a slope approaching that obtained with irradiation under anoxic conditions. The change has been attributed to consumption of oxygen by products of the irradiation. At very high dose rates diffusion of oxygen is unable to make good the deficiency. Evidence in favour of this view is that the breakaway point occurs at higher doses as the oxygen concentration in the medium is raised, shown in Fig. 1. However, Berry et al (8) have recently carried out experiments on mammalian cells which led them to question this view. Using a very intense pulse of X-rays they have observed curves of the type illustrated when the dose was delivered in 7 ns, but with a dose of 50 ns duration the high dose rate curve was almost indistinguishable from that obtained with a dose rate of 100 rad/min. They point out that workers who used pulsed beams from linear accelerators in fact delivered the dose in a large number of very short pulses at the frequency of the RF power supply. In the pulses themselves the dose rates may have been much higher than would have been supposed from the overall duration of the exposure, which was of the order of 1 μs. If the breakaway from the low dose rate curve was conditioned by the diffusion of oxygen it would seem that depletion must have been experienced during each of the very short RF pulses. But in each individual pulse the dose was very much lower than that at which Berry et al. observed the breakaway to take place in their own experiments. They suggest therefore that the phenomenon is a true dose rate effect and arises from the high local concentration of initial products of radiation. Oliver (9) has made an interesting observation on the results of Berry et al. He points out that in these experiments the duration of the irradiation is kept constant while the total dose is varied. Each point on the graph is therefore obtained at a different dose rate, and the curves are not true survival curves in the sense that having arrived at some particular point on the curve one may then proceed by giving a further dose to a later point on the same curve. Each irradiation is carried out at a different dose rate and the true state of affairs might be better represented by a family of curves none of which shows a breakaway of the type seen in Fig. 1. It is not possible at the present time to resolve these questions, but it is clear that in the field of very short, high dose rate irradiations there is great scope for extending our knowledge of the physical and chemical aspects of radiobiology.

Cellular Effects at Conventional Dose Rates

The effect of time on the action of radiation assumes immediate practical importance at the dose rates conventionally used in radiotherapy. Present treatment regimes have been arrived at empirically, and it is possible that improvements could come from a deeper understanding of their radiobiological basis. Here the influence of time is to be accounted for in biological terms. The processes involved are the recovery from sub-lethal damage within the cell, partial synchronisation of the cell cycle, and the changes taking place in the irradiated tissue in general through the agency of selective cell death and reproliferation.

The survival of mammalian cells with increasing dose of radiation of low LET follows a curve such as that shown on the upper part of Fig. 2. There is a shoulder which suggests that some damage must be accumulated before cell killing can take place with full effectiveness. Curve B on Fig. 2 shows what happens when the irradiation is interrupted by a break of several hours. Some recovery of the damage contributing the shoulder has taken place and must be re-established before the curve resumes its old slope, but this is not the only process involved. The dose increment required to compensate for the break in irradiation depends on the length of the interval in the manner shown later in Fig. 5. Curves of this kind have been demonstrated in many systems, and several processes account for their shape. The initial rise is attributed to the recovery of sub-lethal damage which must be reestablished when the treatment is continued. The subsequent fall is attributed to the fact that radiosensitivity varies throughout the cell cycle in a proliferating tissue. Radiation will selectively kill those cells which are most sensitive, leaving those which are in a resistant phase of the cycle. After a time these cells will have moved to a more sensitive phase of the cell cycle and this will offset the extra radiation required to compensate for recovery of sub-lethal damage. If the cells remain well synchronised more than one peak may be detectable. With a longer interval between treatments, proliferation of cells may take place and more dose is required to produce the same survival level simply because there are more cells to kill.

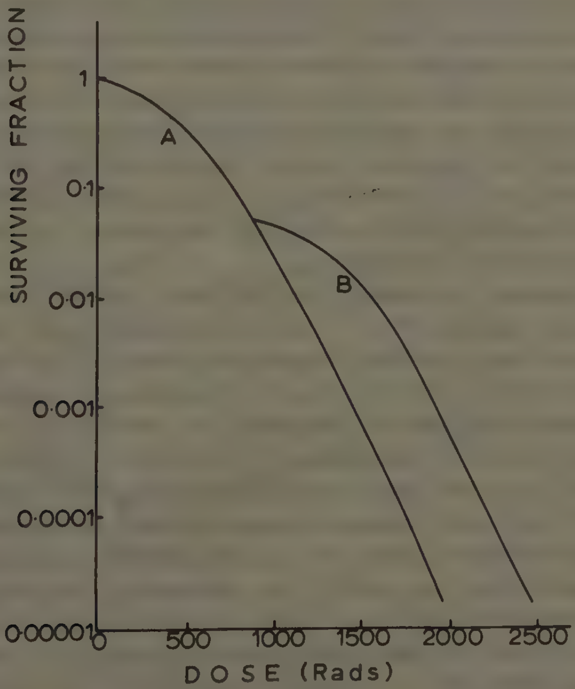


Fig. 2. Survival of mammalian cells vs. X-ray dose (schematic) A—single dose, B—Two doses separated by an interval of several hours.

The effectiveness of radiation shows some dependence on dose rate over the approximate range 1-1000 rads/minute, which is usually ascribed to intracellular repair of sub-lethal damage. Repair proceeds as the radiation is given, and the longer the treatment, the greater is the dose required to produce a given effect. It is usually found that in the range of treatment times common in radiotherapy, approximately 2-30 minutes, there is little dependence of effectiveness upon treatment time, although this is not true when the time is protracted to a day or two as in interstitial treatments. It would hardly be expected that repair would proceed so rapidly as to make the effect of a dose delivered in 10-15 minutes appreciably different from that of a dose delivered in a very much shorter time (though not at the dose rate at which the oxygen depletion effect is observed). However, Hornsey and Alper (10) found a difference of approximately 200 rads (20%) in the LD_{50} for 4 day death in mice irradiated with electrons at 100 rads/minute and 6000 rads/minute respectively. Evidence suggesting that the effect was indeed due to repair was obtained by dividing the treatment at high dose rate into 2 fractions. When the dose at 6000 rads/minute was given in two fractions separated by a 10 minute interval the effect approached that of a single dose at 100 rad/minute. Four day death is associated with intestinal damage, arising from the killing of stem cells in the crypts, and recently Hornsey (personal communication) has confirmed that the survival of individual gut cells as demonstrated by the clone counting technique of Withers and Elkind (11) shows a similar dependence on dose rate.

The effects of X-ray induced cell synchrony have recently been explored extensively by Young and Fowler (12), who have made a comprehensive set of calculations based on Sinclair's results on survival for Chinese hamster cells at various phases of the cell cycle and on the radiation induced division delay (13, 14). It was assumed that the age distribution in the initial asynchronous population was that corresponding to exponential growth without cell loss, and that, where the effect of two doses was considered, repair of intracellular sub-lethal damage was complete before the second dose was given. The effect of a distribution of cell cycle times was also taken into account.

The computations successfully reconstructed the single dose cell survival curve as would be expected, Fig. 3. Two-dose recovery curves were calculated for a number of initial doses with and without a distribution of cell cycle times. A representative set of curves is shown in Fig. 4. In practice, the periodic component of the curves is usually obliterated by the spread of cycle times in the course of a few cell cycles. Fig. 5 shows a comparison of the computer results with experiment. The experimental curve lies below the calculated curve, initially because in the calculation repair of sub-lethal damage is assumed to take place immediately, and the lower survival persists throughout the experimental curve because the dose given was 40 rads greater than that used in the calculations. Subject to this difference the general agreement is good; it is seen that the computed damping out of the oscillation of survival through a distribution of cell cycle times considerably improves the agreement with the practical observations. Two dose survival curves, e.g. Fig. 6, show that if an extrapolation number is derived its value is likely to depend markedly on the interval between doses and also on the survival values

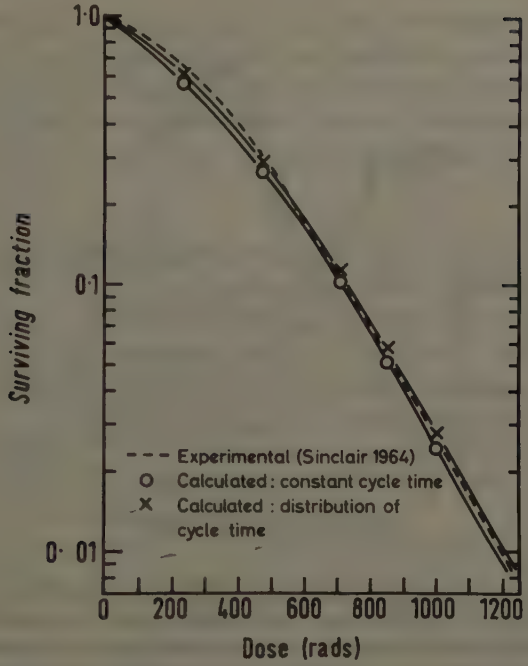


Fig. 3. Comparison between calculated and experimental single dose survival curves for Chinese hamster cells. o, Calculated: constant cycle time; x, calculated, distribution of cycle time;——, experimental (Young and Fowler (12)).

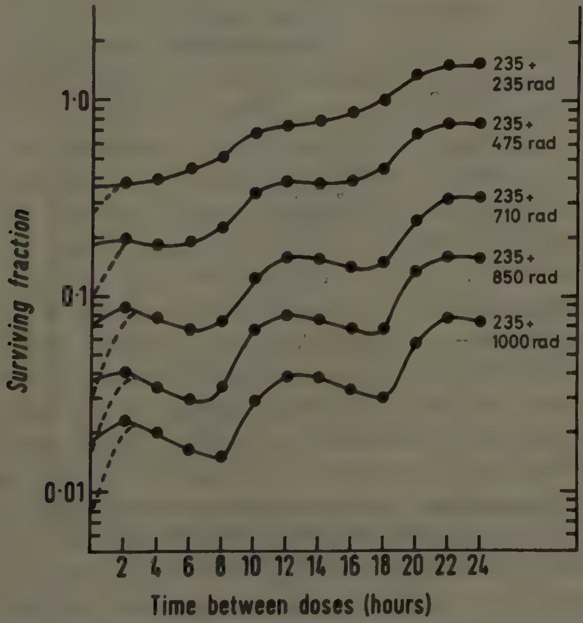


Fig. 4. Calculated two dose recovery curves for a first dose of 235 rads, assuming all cells have the same cell cycle time, and all delay occurs when cells reach G_2.

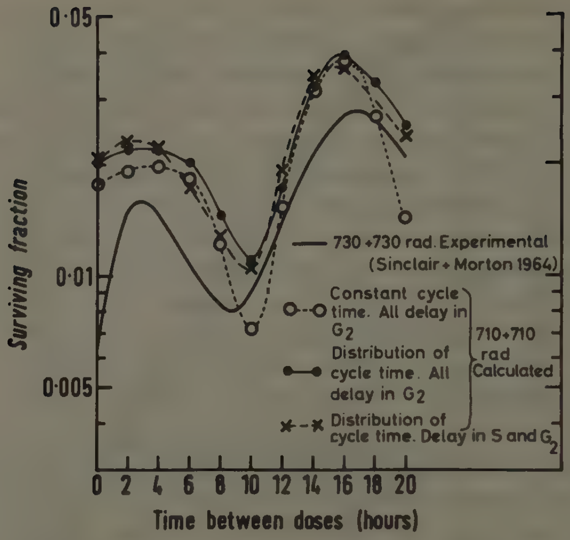


Fig. 5. Comparison of calculated and experimental two dose response curves for Chinese hamster cells treated with X-rays. The unbroken line is an experimental curve for 730 + 130 rads; the broken lines refer to calculated responses for 710 + 710 rads: o, Constant cycle time, all delay in G_2; ● distribution of cycle times, all delay in G_2; x distribution of cycle times, delay in S and G_2 (Young and Fowler (12)).

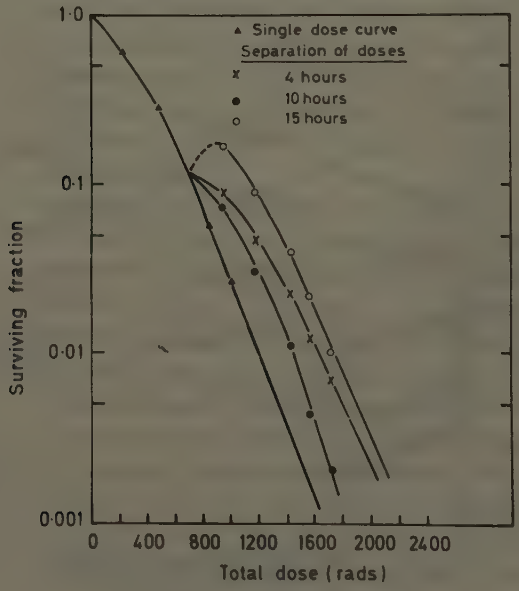


Fig. 6. Calculated two dose survival curves assuming a distribution of cycle times and delay in S and G_2. First dose 710 rads (Young and Fowler (12)).

used. The dependence on initial dose and time between fractions of D_2-D_1 (the extent to which a total dose given as two fractions must exceed a single dose to produce the same effect) was also studied. It was shown that, given a fairly large initial dose and a time interval sufficient to allow the survivors to proceed to a sensitive phase of the cell cycle, negative values of D_2-D_1 can be obtained.

Organized Tissue

The previous section deals with effects on individual cells, such as can be studied in tissue culture or cell suspensions. When organized tissue is irradiated the individual cells are affected not only by the radiation directly, but by the changes that the radiation brings about in the other cells that constitute their environment. Differential proliferation of different types of cell may be important. If, for instance, normal cells divide more rapidly under the stimulus of radiation induced cell depletion than do malignant cells, proliferation between treatments will be advantageous in radiotherapy. In the long term vascular changes and fibrosis may be of the utmost importance.

In the present paper we shall consider, as a particularly important example, the short term effects on oxygenation of the tissue. Many tumours, certainly those with necrotic foci, contain hypoxic cells. Cells in this condition are relatively radio-resistant, and it is probable that in many tumours the dose required to achieve a cure is the dose required to eliminate the hypoxic cells alone, this dose being more than adequate to deal with the well oxygenated cells. Hypoxia results from the balance between tissue demand for oxygen and local blood supply, and radiation may affect the former within a relatively short time. Thomlinson (15) has proposed the model shown in Fig. 7 to illustrate the sequence of events following a dose of radiation. As the tumour grows, its blood supply becomes inadequate and a proportion of hypoxic cells is established which is probably fairly characteristic of the type of tumour. After irradiation, virtually all of the originally well oxygenated cells, because of their high radiosensitivity, sustain damage which ensures that they will die. The cells remaining capable of proliferation are almost all hypoxic. The hypoxic proportion therefore rises to 100 per cent. As the oxygenated cells die, the oxygen demand of the tumour falls and its vascular supply becomes adequate to oxygenate the whole tumour. The previously hypoxic cells now become well oxygenated and the hypoxic proportion drops towards zero. As the surviving cells proliferate, the vasculature once more becomes inadequate to meet their oxygen demands; and the hypoxic proportion reverts to its original value determined by the structure and metabolic characteristics of the tumour. Howes (16) has recently made a contribution to the experimental evidence in favour of this model. He has studied the X-ray dose required to control transplanted C_3H mouse mammary tumours, control being indicated by the absence of recurrence by 150 days. The tumors were irradiated at a mean diameter of 6 ± 0.5 mm and only those with approximately spherical form, growing at a rate between 0.2 and 0.4 mm per day were used. They were irradiated either clamped to produce hypoxia, or under air, or high pressure oxygen breathing conditions. Treatments were given either as single doses or as split doses with various intervals

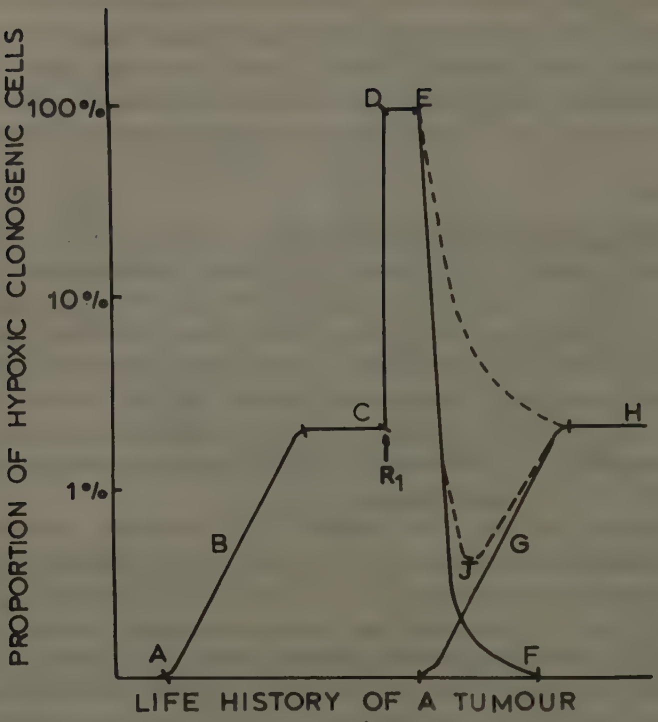


Fig. 7. Variation in the proportion of hypoxic clonogenic cells in a tumour. A dose of radiation is given at R_1. A second dose would be most effective at the minimum, R_2. The depth of the minimum depends on the extent to which the oxygen demand is reduced by cell death, before it is increased again by proliferation of survivors (Thomlinson (15)).

following a priming dose of 1500 rads. The experiments yield values of the TCD_{50}, the dose required to control 50 percent of the tumours, under the various treatments and are in general agreement with Thomlinson's model. A rise in hypoxic proportion is indicated by a movement of the TCD_{50} in air or oxygen breathing animals towards the value obtained for a hypoxic (clamped) tumour. Such a rise was apparent 1 hour after the priming dose, but by 6 hours there was already evidence of re-oxygenation. Given reasonable assumptions about the oxygen enhancement ratio and the reciprocal slope of the hypoxic survival curve, it is possible to make an estimate of the absolute number of hypoxic cells in the tumour. Fig. 8 presents the results of the investigation in this form. It is seen that the number of hypoxic cells is minimal between 2 and 3 days after the first dose of radiation. A second dose would therefore be most effective given at this time. This type of investigation is, of course, practicable only with a tumour of well defined characteristics growing in an inbred strain of animals. If means could be devised of predicting the variation of anoxic proportion in an individual tumour in a human subject it would be of inestimable value in radiotherapy.

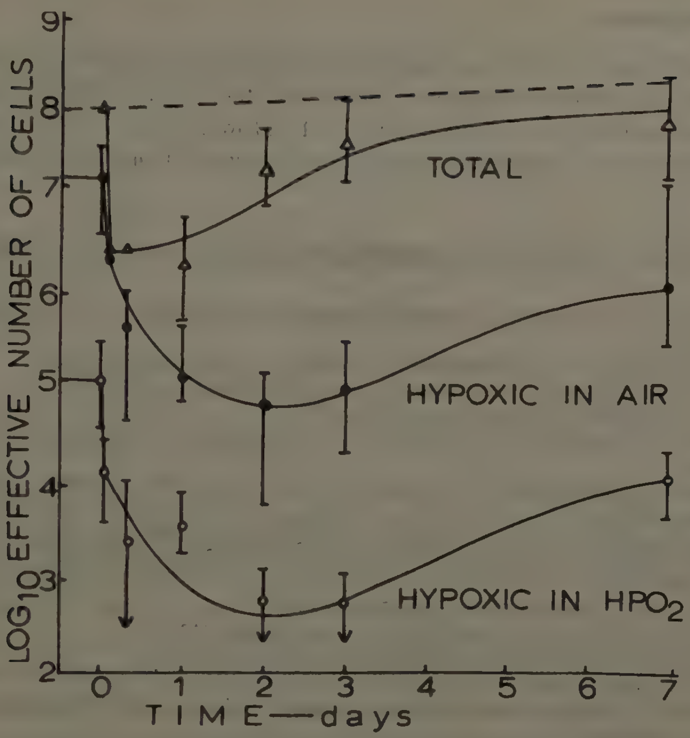


Fig. 8. Graph showing estimated values of absolute numbers of tumour cells. The upper dashed line represents the mean growth of unirradiated tumours (volume doubling time six days). The open triangles represent values of the total number of viable tumour cells at various times after 1,500 rads, assuming the tumour initially contains 10^8 viable cells. The full circles represent values of the effective number of hypoxic cells which exist during air breathing, and the open circles represent values for high pressure oxygen breathing. The vertical bars indicate 95 per cent confidence limits obtained from corresponding errors in TCD_{50} values.

Very Low Dose Rates

It is beyond the scope of this paper to deal with what is probably the most important range of time scale, the effect of protracting irradiation over a large part of the life span of the individual. As radiation becomes increasingly part of the environment it is necessary to know as much as possible of the risks involved in its use. Much painstaking work has been done in this field. The only experiments of unquestionable validity are those which are in fact made over a very long period, and the most relevant information, that in human subjects, is notoriously difficult to collect. On the other hand, results can be more rapidly obtained by making studies at high dose rates and extrapolating to low dose rates. A most valuable aspect of radiobiology is that it may one day lead to a sufficient understanding of the fundamentals to allow this extrapolation to be made with confidence.

Acknowledgements

It is a pleasure to acknowledge many helpful discussions with Professor J. F. Fowler and many other colleagues. Thanks are due to Dr. E. R. Epp and his colleagues, and to Prof. J. F. Fowler, Dr. A. E. Howes, Miss J. M. Young and Dr. R. H. Thomlinson for permission to reproduce diagrams from their publications.

REFERENCES

1. Barendsen, G. W., Walter, H. M. D., Fowler, J. F. and Bewley, D. K. (1963). *Radiat. Res.,* **18,** 106.
2. Epp, E. R., Weiss, H. and Santomasso, A. (1968). *Radiat. Res.,* **34,** 320.
3. Dewey, D. C. and Boag, J. W. (1959). *Nature (Lond.),* **183,** 1450.
4. Town, C. D. (1967). *Nature (Lond.),* **215,** 847.
5. Todd, P. W., Winchell, H. S., Feole, J. M. and Joses, G. (1967). *Radiat. Res.,* **13,** 644 (abstract).
6. Kennon, W., Malkinson, F. D., Skeggs, L. S. and Griem, M. L. (1968). *Radiat. Res.,* **35,** 569 (abstract).
7. Phillips, T. L. and Worsnop, R. B. (1968). *Radiat. Res.,* **35,** 564 (abstract).
8. Berry, R. J., Hall, E. J., Foster, D. W., Storr, J. H. and Goodman, M. J. (1965). *Brit. J. Radiol.,* **42,** 102.
9. Oliver, R. (1969). *Brit. J. Radiol.,* **42,** 231.
10. Hornsey, S. and Alper, T. (1966). *Nature (Lond.),* **210,** 212.
11. Withers, H. R. and Elkind, M. M. (1969). *Radiat. Res.,* **38,** 398.
12. Young, J. M. and Fowler, J. F. (1969). *Cell and Tissue Kinetics,* **2,** 95.
13. Sinclair, W. K. (1967). Radiation Research, ed. G. Silini, p. 607 (North Holland Publishing Co., Amsterdam).
14. Sinclair. W. K. (1968). Biophysical Aspects of Radiation Quality, Second Panel Report, p. 39 (I.A.E.A., Vienna).
15. Thomlinson, R. H. (1968). *Front. Rad. Ther. Onc.,* **3,** ed. J. M. Vaeth, p. 109 (Karger, Basel).
16. Howes, A. E. (1969). *Brit. J. Radiol.,* **42,** 441.
17. Rossi, H. H. (1967). Adv. in Biol. and Med. Phys., Vol 2, p. 27 (Academic Press, New York).

STATISTICAL ASPECTS OF RADIATION ACTION

Albrecht M. Kellerer

Early radiation biology had a particular attraction to the biologist as well as the physicist. The kinetics of radiation action seemed to indicate that in the living cell a single atomic or molecular random event can express itself in a macroscopic result. This induced high expectations into the use of ionizing radiation as an instrument to probe the structure of living matter. Today, some fifty years later, these expectations have been somewhat dimmed. Still the questions asked in the early days of radiation biology have lost little of their interest. It may therefore be useful to point out the directions which radiation biology has been taking after the setbacks encountered by some of the old ideas.

First some of the concepts may be recalled which have shaped quantitative radiobiology in the past. Viruses and certain bacteria show an exponential decrease of survival probability with radiation dose. The exponential relation is characterized by the fact that equal dose increments decrease the fraction of survivors by equal ratios. In analogy to the kinetics of radioactive decay this has been taken as an indication that the radiation damage is not due to continuous accumulation of energy absorption. Instead it was assumed that the effect is brought about in a single critical event. These hypothetical events were not identified by Dessauer (1), who tentatively spoke of 'point heat'. Crowther (2), however, who simultaneously developed the statistical concept of radiation action proposed that the critical events were in fact single ionizations.

The single hit concept was the starting point for a formalism which is based on the assumption that the statistical fluctuations of energy deposition and not the reaction kinetics of the cell determine the survival curve. It was assumed that this is true not only for exponential dose effect curves but also for 'sigmoidal' curves which were considered to be the expression of damage accumulated in a certain number of statistically independent critical random events.

The statistical models of radiation biology have been discussed in a vast literature of their own. Biologists have always been hesitant to accept the highly formal analysis of dose effect relation which has been developed in target theory (3). But it was only in recent years that the weak points in the conventional analysis have been pinpointed, and that alternative approaches have been indicated (4). There are two essential aspects: First, the conventional formalism is mathematically inconsistent. Secondly, it is wrong to assume that all random factors which express themselves in the survival curve are connected with the fluctuations of energy deposition; additional stochastic factors are equally important.

In order to outline the implications of these facts there will be first some general remarks on the analysis of the random factors involved in radiation action. Then the special problem of the statistics of energy absorption will be discussed. Finally some of the random factors characteristic for the biological processes will be dealt with.

The dose response relation as an expression of statistical processes.

The conventional evaluation of dose effect curves has been based on the assumption that the biological effect, e.g. cell inactivation, is due either to a single energy absorption event or to a succession of random events of energy absorption in the cell. One obtains characteristic dependencies of inactivation probability on dose if one assumes different numbers of events for the critical threshold. There have been various methods to distinguish the resulting curves and to infer the critical threshold values. The mathematical techniques employed have not always been the most efficient ones, but the basic ideas have been the same.

One may summarize the reasoning by the statement that, if the critical number of statistically independent events is n, then one obtains a so-called gamma-distribution of order n as reaction curve. The inactivation probability, $P(D)$, as a function of absorbed dose is then given by:

$$P(D) = \int_0^D ke^{-kD} \frac{(kD)^{n-1}}{(n-1)!} dD \qquad (1)$$

where k is the mean number of critical events per unit dose.

Equation (1) is more commonly given in the form:

$$P(D) = \sum_{\nu=0}^{n-1} e^{-kD} \frac{(kD)^{\nu}}{\nu!}$$

It is easy to show that both forms are equivalent. Equation (1) has the advantage that it clearly shows that the differential distribution, $p(D)$, of the inactivation dose is indeed a gamma-distribution:

$$p(D) = \frac{dP(D)}{dD} = ke^{-kD} \frac{(kD)^{n-1}}{(n-1)!} \qquad (3)$$

Fig. 1 represents the inactivation curves. Fig. 2 represents the corresponding differential distributions of inactivation dose. Probability theory offers simple ways to distinguish these curves and to infer the characteristic index n. The easiest method is a comparison between the mean inactivation dose and its variance. If few random

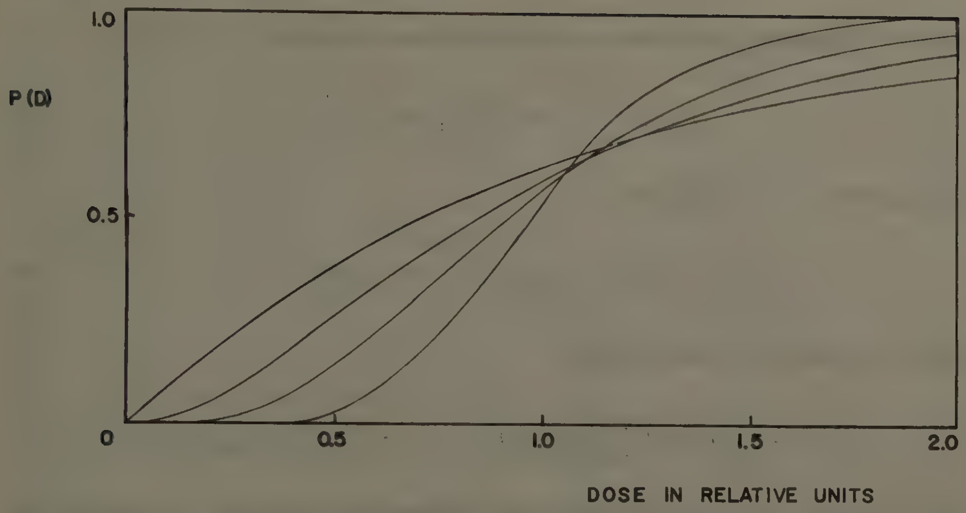


Fig. 1. Reaction curves according to equ. (1) or (2). The curves are integral gamma-distributions of order 1, 2, 4, and 10. The dose is in arbitrary units.

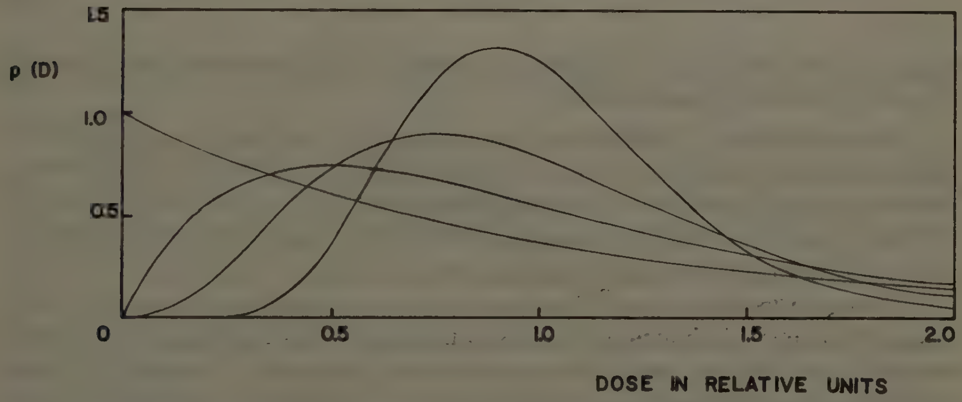


Fig. 2. Probability densities of inactivation dose according to equ. (3). The curves are the derivatives of the curves in fig. (1), i.e. they are gamma-distributions of order 1, 2, 4, and 10.

events are sufficient to bring about the effect, the random nature of the process is strongly expressed, and the response curve instead of being a step-like function must be fairly washed out. If a great number of events is involved in cell inactivation, the random fluctuations are less expressed and one may obtain a curve which is rather steep. Thus the standard deviation, σ, of the inactivation dose as compared to the mean inactivation dose, $\overline{D}$, indicates the influence of the random factors involved in the process. The mean inactivation dose is defined as:

$$\overline{D} = \int_0^\infty D\, p(D)\, dD \qquad (4)$$

which in the special case of the gamma-distribution leads to:

$$\bar{D} = k \int_0^\infty D\, e^{-kD} \frac{(kD)^{n-1}}{(n-1)!} dD = \frac{n}{k} \tag{5}$$

The variance, σ^2, is defined as:

$$\sigma^2 = \int_0^\infty (D - \bar{D})^2\, p(D)\, dD = \int_0^\infty D^2\, p(D)\, dD - \bar{D}^2 \tag{6}$$

and with equation (3) one obtains:

$$\sigma^2 = k \int_0^\infty D^2\, e^{-kD} \frac{(kD)^{n-1}}{(n-1)!} dD - \frac{n^2}{k^2} = \frac{n(n+1)}{k^2} - \frac{n^2}{k^2} = \frac{n}{k^2} \tag{7}$$

Thus the ratio D^2/σ^2 which measures the *steepness* of the curve is equal to the critical number of events:

$$\frac{\bar{D}^2}{\sigma^2} = n \tag{8}$$

It is remarkable that only the first two moments of a dose-response curve have to be known to obtain the parameter n.

The mean value, $\bar{D}$, depends on the cross-section constant k (see equ (5)). The absolute position of the curves is therefore irrelevant to the statistical analysis. The curves in fig. 1 and 2 are arbitrarily normalized to the same mean value. The width of the curves indicates how the relative deviations from the mean inactivation dose $\bar{D}$ decrease with increasing value of n.

These considerations have been dealt with in some detail because they have actually been applied, though not always in the most consistent mathematical form, to a number of different fields. The application to the action of ionizing radiation has already been mentioned; an excellent survey of this application has been given by Zimmer (5). Independently the formalism has been used to derive the critical number of light quanta to evoke a light sensation in the human eye (6,7,8,9). If one plots observation frequency as a function of the intensity of a short light flash one should see the influence of the statistical fluctuations of the number of photons striking the retina. By analysing this curve the threshold numbers have been determined by various workers, who have found various values for the critical numbers; an example is given in figure 3. A third application has been in pharmacology where the possibility has been discussed that toxic effects on cellular units are the result of the action of a small number of molecules of the pharmacon (10,11,12). One has seriously considered this even in such an extreme case as the killing of the large alga *Nitella* by copper-chloride. It is quite interesting to follow the discussion of this problem for example in Clark's chapter in the handbook of pharmacology (13). Apparently quite similar discussions have been aroused by the respective application in radiobiology, physiology, and pharmacology, but the

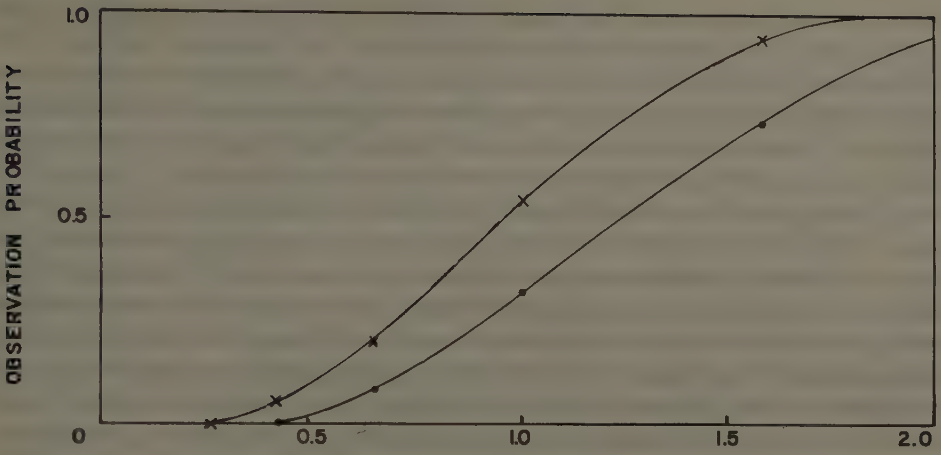


Fig. 3. Observation frequency as function of the intensity of short light flashes. The curves are from data presented by HECHT et al (7).

workers in these different fields seem to have been largely unaware of the parallel approaches.

The weakness of the formal interpretation of the dose-response relations is that it is based on an idealized model. In general—some of the light quanta experiments may be an exception—one had been well aware of this limitation (3,5). Specifically the biological variability of the exposed objects and the possible differences in the hypothetical critical events have been discussed. But more or less tacitly it was assumed that the sum total of all the deviations from the idealized model might cancel out, so that the analysis, though blurred by additional parameters, might still be essentially right. This assumption would imply that if for example the attention of the observer in the light quanta experiments fluctuates this would on the average cancel out. Similarly if the units of an irradiated population responded differently the compound curve would have a shape which corresponds to the mean of the individual hit numbers and threshold values.

This confidence in the mathematical formalism proved to be fundamentally wrong. It has been shown that, far from averaging out, all deviations from the idealized model work in one direction. The general effect is one of broadening the response curve i.e. increasing its relative variance. In other words, the deviations suggest a critical number which is too low. This can be illustrated by the two curves of fig. 3. The parameters n deduced from these curves are 6 and 7. If one pools the data from the two experiments and calculates the parameter n for the averaged curve one does not obtain $n = 6.5$. Instead one finds a value of approximately 3. Thus the relative variance is sharply increased by averaging the two experiments. It is interesting to note that the increase is appreciable even in this case where the difference in the reaction of the test objects is moderate. If one studies the response of mammalian cells to ionizing radiation in the various phases of their generation cycle one finds differences which are much more expressed. The relative variance of the dose effect relation must then be largely due to the fluctuations in cellular sensitivity. The fact is indeed brought out in studies with synchronized cell cultures. But even with a well synchronized population one can not assume that σ^2 solely reflects fluctuations in energy deposition.

Therefore all one can formally deduce from a dose-response relation is a lower limit for the number of interacting random events. This is true in the light quanta experiments for the threshold of vision, it is clearly true in pharmacological reaction curves, and it applies to radiation biology as well. Formally one can always deduce the order of the reaction kinetics which expresses itself in the dose-response curve. But this is in fact a purely formal index influenced by the full complexity of the reaction of the biological object to an irradiation. It is rather important to keep this in mind as it has too often been overlooked. In the determination of cross-sections one is well aware of the fact that these are in general formal indices which represent only a lower limit for the true physical cross-section. One should be equally aware that the same is true for the reaction order of a survival curve, evaluated on the basis of whatever particular model.

There is little reason to believe that the interplay of fluctuations of energy loss, of biological variability, and of statistical behaviour of the cell can be adequately described either by equation (1) or any of its variations employed in target theory. Curve fitting formulae with or without target theory interpretation have brought little progress to radiation biology. Application of the theory of stochastic processes seems more promising and has already lead to some useful results (4). One may just note one particular finding which is connected to the discussion on the moments of the dose effect curve given earlier in this paragraph. It has been proved in full generality that for an experimentally determined relative variance $V = \sigma^2/\bar{D}^2$ the mean number of statistically independent energy absorption events involved in the effect can not be smaller than V^{-1}. This is a result not without interest. From certain experiments on the x-ray inactivation of mammalian cells in S-phase one can for example deduce that more than 4 energy absorption events (passages of ionizing particles) must on the average be involved in the inactivation process. This is a valuable conclusion, but it is, of course, a rather limited statement as compared to the original idea of deriving an actual 'event number'. The example of the quantum vision experiments makes this particularly clear. The experiment has been designed to show how *few* quanta are needed to evoke a light perception. Formal analysis, however, can only indicate how *many* quanta must be at least involved in the process, the actual number being perhaps much higher. This means that formal analysis has to be supported by additional physical evidence if one wants to bracket the number of interacting random events more closely. In the case of the light quanta experiments one simply has to determine the actual mean number of photons which hit the retina (7,8,9). With ionizing radiations the physical measurements are more complex. In fact they form a discipline of their own for which the name microdosimetry has come into use.

The new approach to the energy absorption problem

A realistic treatment of the statistics of energy absorption has to be based on the analysis of the microscopic spatial patterns of energy deposition with different kinds of ionizing radiation. These patterns are too complicated to be adequately represented by either a Poissonian distribution of statistically independent ionizations or by straight, continuous tracks of charged particles as suggested by the theory of linear energy transfer (LET). A theoretical analysis (4,14) of these patterns involves various factors. Some of these are: track segment length in a sensitive volume or in a region of interaction of free radicals, curled shape of the particle track, varying stopping power and, last but not least, the discontinuous nature of energy deposition along the track. The theoretical analysis of the interplay of these factors has to rely on large scale electronic computing, it must also be based on quantum mechanical data on collision cross sections which are as yet insufficient. It is for this reason that H. H. Rossi and co-workers used a more direct experimental approach when they introduced the concepts of microdosimetry (15,16). They simulated the processes of energy deposition in tissue by employing spherical proportional counters filled with tissue equivalent gas and surrounded by tissue equivalent walls, in order to directly determine the probability distribution of the energy deposited in microscopic volumes. This experimental analysis of the patterns of energy absorption has been performed for various kinds of ionizing radiations and it has lead to very useful conclusions. The resulting probability distributions of specific energy—or, in more naive terminology, local dose—are limited to regions not much smaller than one micrometer. But there seems to be no reason why these distributions cannot also be derived, theoretically or experimentally, for smaller regions. Moreover it appears that in one of the systems of foremost interest to radiation biology, the mammalian cell, the biologically relevant interactions of energy depositions cover distances of the order of magnitude of a micrometer.

The case of a tissue sphere of 0.5 μm diameter exposed to ^{60}Co γ-radiation is taken as an example in fig. 4 to illustrate the probability distributions of specific energy, z, and their dependence on dose D. Without going into the actual definitions of the basic quantities and probability distributions in microdosimetry one may point out a few approaches in order to characterize the possible applications to radiation biology.

One of the most straightforward applications is the one which is used to show that in certain cases exponential survival curves are indeed due to single events. An outstanding example is the exponential survival curve derived for mouse spermatogonia in vivo with neutron irradiation (17). Microdosimetry data show that at the 37% survival the mean number of events in the whole cell is near to one. That indicates that with this particular radiation and in this special cellular system exponential survival is indeed an expression of single hit inactivation; it further indicates that a heavy charged particle kills the cell it passes even if it does not penetrate its nucleus. One might still assume that cell-killing in these *in vivo* exper-

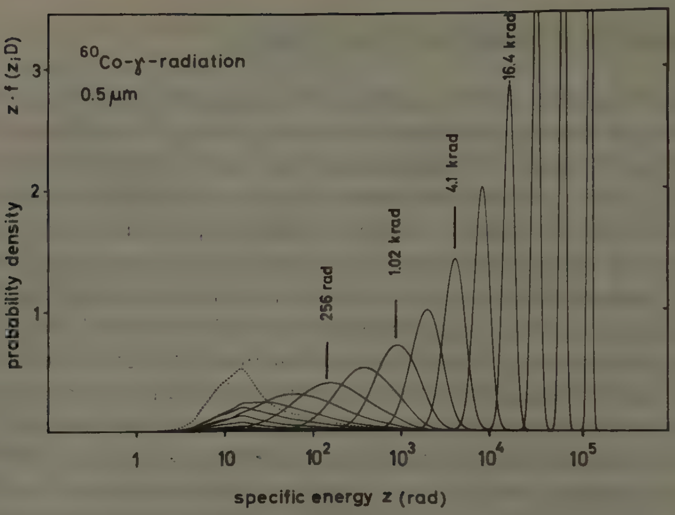


Fig. 4. An example of the distributions of specific energy, z. The curves correspond to the irradiation of a tissue sphere of 0.5 μm diameter with ^{60}Co-γ radiation. The dose, D, is given as parameter. For details see (14).

iments is also influenced by humoral factors in the tissue in which the spermatogonia are imbedded. Thus the case for a single hit process is not established beyond all doubt. What microdosimetry data tell, however, is that cell killing in this instance is certainly not due to a multi-absorption event process inside the cell.

One must, of course, be careful not to blindly extrapolate these findings and to conclude that all exponential curves are due to single-absorption event action. This point will be somewhat further discussed in the next section.

A second type of application is somewhat more indirect. It involves the assumption of a well-defined sensitive site in the biological object. Therefore it can only be accepted with the same reservations which one may hold regarding this assumption. An example is the analysis of the mutation frequencies in maize produced by neutrons of different energies (18). The model chosen by Rossi and Smith for evaluation of these experiments is that of a sensitive volume with a certain threshold of absorbed energy for induction of this specific mutation. A comparision of the probability distributions of energy absorption for the different neutron energies with the observed mutation frequencies leads to the conclusion that for a region of diameter 0.3 μm and only for this diameter one obtains one and the same critical threshold regardless of radiation quality. There is support for these results in findings which Sparrow, using a different experimental endpoint, has obtained in recent years on various plant species (19). His results suggest that an irradiated plant is killed if a critical amount of energy is deposited in its chromosomes. The notion that the chromosomes are the gross-sensitive structures in radiation effects on plants conforms with the assumption that the critical distances of interaction of energy deposition are of the order of magnitude of a micrometer.

The probability distributions of local energy density can, by the way, also be used to demonstrate that the simple model of a spherical sensitive volume is to be rejected in some cases. An example are studies of lens opacification (20). It is noteworthy that here the analysis leads to the concept of two adjacent sensitive sites or to the assumption that biological effects depend not only on a certain critical energy threshold but also on a certain spatial extension of the energy deposition. This idea is in agreement with models Neary has proposed in recent years (21) to explain the induction of chromosome aberrations in *Tradescantia* by radiation of different ionization density.

Microdosimetry may also be applied in a more general fashion without actual reference to any particular model of radiation action. As was pointed out earlier, the dose response function is not merely a reflection of the energy absorption processes. Other statistical factors may as well be involved. The objects in the irradiated population differ or are in different phases of the cell-cycle. Moreover the reaction chain from the primary steps of energy deposition to the formation of the initial lesion and to the final manifestation in a chromosome aberration or in a mitotic failure involves stochastic elements. This is specifically so if one deals with sigmoidal dose response functions which express a continuous or stepwise accumulation of damage. In these cases it may be appropriate not to try to explain the detailed shape of the curve. Instead one may analyze the relative variance of the observed survival curves on the basis of microdosimetry data. The variance of the dose response curve is only partly due to the fluctuations of energy absorption. In general one cannot split up the observed variance according to the different stochastic factors involved, thus one may merely conclude that the variance due to the energy absorption statistics is equal or smaller than the one which has been experimentally observed. The relative variance of the local energy density distributions is largest for the smallest volumes. It is therefore possible to derive a lower limit for the diameter of the sensitive sites below which the fluctuations of energy absorption alone would exceed the total relative variance of the survival curve (4). From inactivation experiments on mammalian cells with Cobalt gamma radiation and x-rays one concludes in this way that the gross sensitive volume cannot be smaller than 1 μm. Otherwise the statistical spread due to the fluctuations of energy absorption would be greater than the observed variance of the dose response curve. This is a finding which shows that microdosimetry can be applied where the classical formalism for the analysis of sigmoidal dose response curves fails.

In order to illustrate this fig. 5 gives two examples of inactivation curves obtained on cultures of isolated mammalian cells. Fig. 6 represents the curves which result if one asks for the probability to exceed a certain value of z at a certain dose. These curves refer to a spherical tissue region of a diameter of 1 μm and to x-rays. The inactivation curves are similar to the distributions of dose, D, necessary to reach a local dose z of about 300 rad in the 1 μm-sphere. This does not imply that one deals in fact with a sensitive site of 1 μm and with a well-defined threshold. The sensitive regions in the cell may be, and probably are, of complicated structure, they may also consist of many separate sites. The comparison of fig. 5 and fig. 6 merely shows that the gross sensitive area cannot

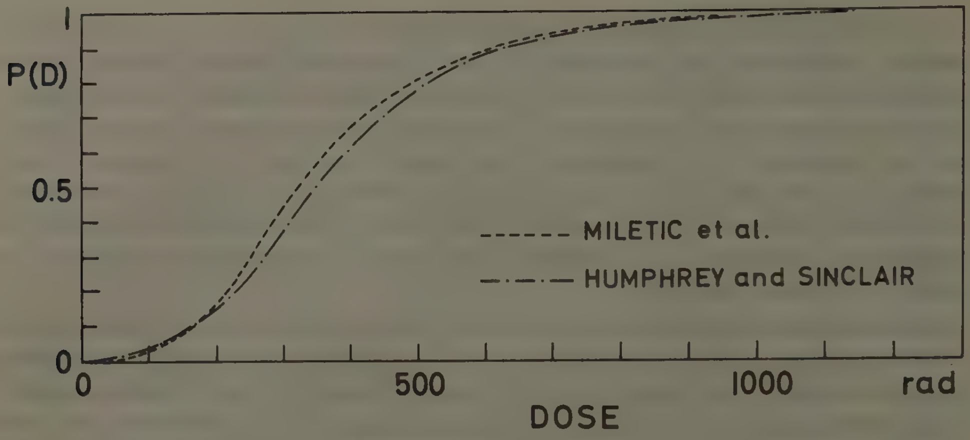


Fig. 5. Two examples of an inactivation curve of mammalian cells with x-rays. No confidence limits are given here; for an estimate of the uncertainty in the curve shape see (22) and (23).

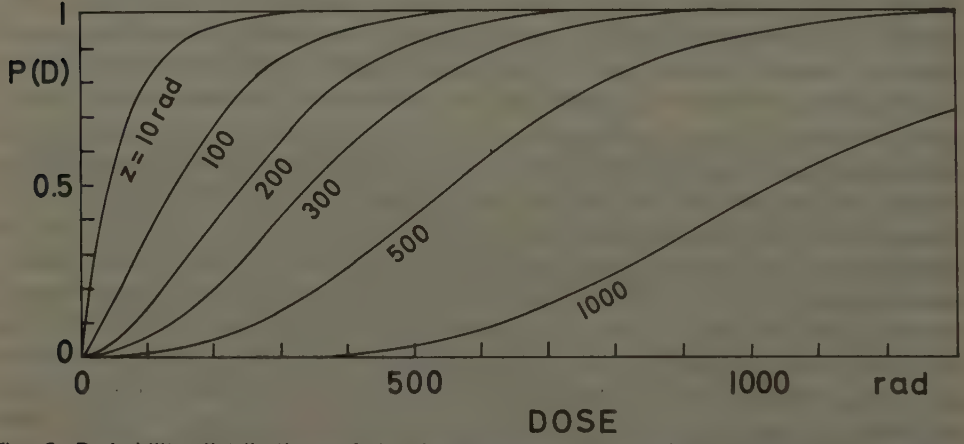


Fig. 6. Probability distributions of the dose necessary to reach certain limits of specific energy, z, in a 1 μm diameter tissue-sphere. The threshold values of z are given as parameters. The curves are based on calculated distributions for 200 keV x-rays (4).

have a diameter less than 1μm. The sensitive structures must be dispersed in the cell in such a way that they cannot be included in a sphere of diameter less than 1 μm.

Finally microdosimetry will certainly be of great importance in a more quantitative treatment of radiation chemistry problems. The concept of LET is severely limited even if one deals with the comparatively simple system of a homogeneous aqueous solution, and it has to be substituted by the analysis of the actual distribution of relative distances between tracks and between energy depositions along and around the tracks. Straggling, i.e. the discontinuous energy deposition along the particle track, is in many cases more important than LET (24). This accounts for the fact that the actual mean values of local energy concentration in regions of roughly 1 μm diameter are very nearly proportional to the empirically determined values of the so-called quality factor while mean values of LET are not (25). Microdosimetry data can certainly lead to a much better understanding of RBE than the LET concept or any refined model based on it.

Randomness due to physiological processes

The cases in which a dose response curve reflects the statistical fluctuations of energy deposition are most characteristic for radiobiology. For this among other reasons, studies with densely ionizing radiations, where one usually obtains exponential dose effect relations, meet particular interest in the field. Conclusions from these experiments must, however, not be generalized to sigmoidal dose response curves without further evidence. It has been mentioned in the preceding paragraph that when a sphere of a diameter of 1 μm is exposed to x-rays, the fluctuations of energy deposition are large enough to explain all of the observed variance of the dose effect relation in the inactivation of mammalian cell cultures. But the fluctuations become much less expressed when one deals with larger volumes. If the whole nucleus of the cell were equally sensitive the statistical variations of energy deposition with x-rays would be rather insignificant. The variance of the dose effect response would then be entirely due to other stochastic factors.

A comparison of the experiments on unsynchronized and synchronized cell cultures shows that biological variability due to different sensitivity of the cell in the various phases of its generation cycle plays an important role. Work with synchronized cultures can at least partly eliminate this factor. It should, however, be noted that the application of the conventional formalism in these experiments may be particularly misleading. It has become more and more common in radiobiology to represent dose effect relations against a logarithmic scale of survival probability. Historically this has developed from certain target theory models. These models have been given up by now. But the use of the logarithmic scale of survival rate has been found practical since it stresses the low survival part of the dose effect curve, and this part is easiest to determine experimentally. For this reason the target theory plot of survival curves is still preferred. It is also preferred because the low survival part of the dose effect relation is specifically relevant, or is thought to be specifically relevant, to radiotherapy. While these may be perfectly good reasons to adhere to the conventional representation of dose effect curves, one should still be aware of the fact that this representation leads to a somewhat biased view of cellular radiation response. Quantities, as for example the so-called extrapolation number, are strongly influenced by small subpopulations of cells with different response. A mathematical analysis based on the moments of the dose response is therefore necessary in order to obtain improved quantitative data on the variations of cellular sensitivity and response characteristics within the generation cycle.

In an ideally synchronized cell culture of a perfectly homogeneous strain one still had to expect stochastic factors in the cellular response. It has been shown that even exponential survival curves may be explained on the basis of the stochastic behaviour of the cell (26,4). In fact, it is characteristic for a multi-component system that its overall performance rate decreases exponentially if the small breakdown probability of its individual components is increased linearly. Mitosis, specifically in an *in vitro* system, is a critical phase in cell life and as such susceptible to an inherent lability. The spontaneous breakdown probability expresses itself in the limited plating efficiency in cell culturing and in occasional unsuccessful divisions throughout the development of a clone.

Till et al. (27) have introduced the concept that the survival curve may be an expression of the random pattern of successful and unsuccessful divisions within pedigrees of the irradiated cells. They performed Monte Carlo calculations and demonstrated the influence of this factor. The problem can also be solved analytically, it is in fact a well-known example of elementary probability theory, the so-called "gambler's ruin" problem (4). The solutions of this problem show that even with an exponentially decreasing division probability one obtains sigmoidal dose dependence for the probability to reach a certain clone size. This is illustrated by the curves of fig. 7, and it supports the observation that the shape of the survival curve indeed depends on the choice of the experimental end point. From observations on the pedigrees of cells it is obvious that the stochastic aspect of clone development must influence the dose effect relations. Experimental data on mammalian cells are as yet scarce, but detailed pedigrees obtained for yeast cells exposed to UV-radiation (28) agree well with the simple model that the division probability is reduced exponentially with dose (29).

Although extensive Monte Carlo studies have been performed on the basis of the stochastic approach to cellular kinetics (30), the results are still of qualitative rather than quantitative nature. It is not known at present how the observed variance of the various survival curves of mammalian cell systems splits up into the contributions of the different statistical factors involved.

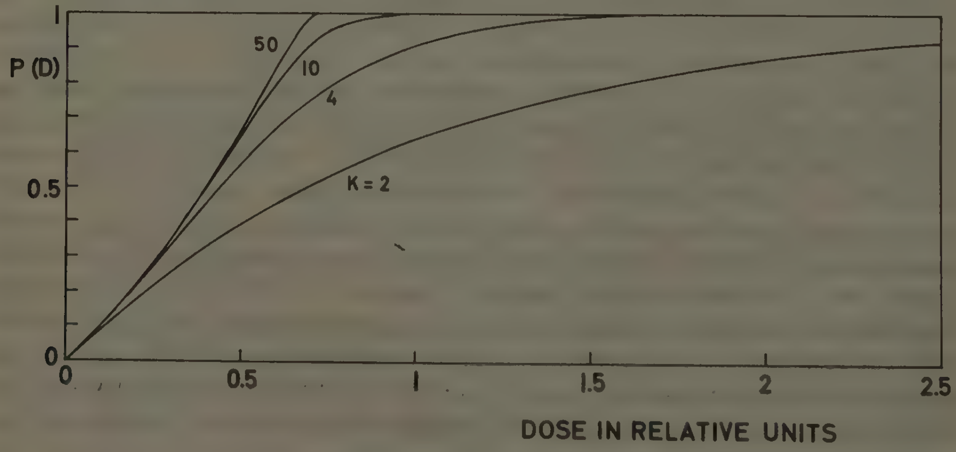


Fig. 7. The probability for failure of a single cell to develop into a clone of at least K cells. The probability is plotted as a function of dose, under the assumption that the division probabilities after irradiation decrease exponentially with dose. For a detailed discussion see (4).

A few concluding remarks on the meaning of the initial slope of the survival curve may demonstrate the ambivalence in possible interpretations. Whatever the detailed shape of the survival curve may be, the initial slope is in general looked upon as a direct indication that the biological effect can be brought about by a single act of energy deposition. This is one of the simplest and most widely accepted assertions concerning the dose effect relation. It can be generalized according to the fact that a dose response curve may be expressed as a power series at zero dose. If $P(D)$ is the fraction of cells which form no visible clone one obtains the following dependence on dose D:

$$P(D) = a_0 + a_1 \cdot D + a_2 \cdot D^2 + \ldots$$

Then a_o is the spontaneous failure rate, a_1 represents the initial slope and is interpreted as single hit probability per unit dose, a_2 is assumed to be the probability that at unit dose two 'sublethal' events occur and interact to inactivate the cell.

A good point can be made for this interpretation. It is in general not easy to obtain accurate figures for the initial slope a_1. Particularly the corrections for cellular multiplicity as they are generally employed may distort the actual values. But with the help of some additional information as for example provided by Hall and Bedford in their cell inactivation experiments with very low dose rates (31) one can at least obtain estimated values of a_1. Microdosimetry data show that even with sparsely ionizing radiation there is always a certain probability that high local energy densities are produced in the passage of an ionizing particle. It is then not surprising that there should be a single hit component in the survival curve.

Nevertheless it is unjustified to blindly identify numerical estimates of a_1 and a_2 with event probabilities. Reasonably accurate survival probabilities can at present not be derived with x-ray doses below 50 rad. At a dose of 50 rads however a cell is still penetrated by several hundred ionizing particles. Even the nucleus is passed by roughly a hundred charged particles at this dose. One can therefore not exclude that in this dose range the division probabilities after irradiation are well-defined functions of dose without much influence of the statistical fluctuations. If one assumes that the breakdown probabilities in the first and second generation are q_1 and q_2 D and if higher order terms and cell deaths in later generations are not considered, one obtains the inactivation probability:

$$P(D) = q_1 \ D + q_2^2 \ D^2$$

Thus the coefficient in the linear term is interpreted as breakdown probability in the first post-irradiation generation, while the square term represents failure in the second generation. Again the linear term corresponds to the probability that the effect is brought about in a single random event, while the square term reflects the probability for the interaction of two random events, namely the failure of both daughter cells in the second post-irradiation generation. This is yet another example that one and the same formal interpretation of the dose effect relation can be given completely different meaning. It is for this reason that formal analysis must be closely supported by microdosimetric data and by studies on cellular kinetics and its radiation induced disturbances.

Acknowledgement

The author is much indebted to Dr. Otto Hug, who developed many of the ideas expressed in this paper.

REFERENCES

1. Dessauer, F., Über einige Wirkungen von Strahlen *I. F, Z. Physik* **12,** 38 (1922).
2. Crowther, J. A., Some considerations relative to the action of X-rays on tissue cells. *Proc. Roy. Soc.* (London) **B96,** 207 (1924).
3. Timofeeff-Ressovsky, N. M., u. K. G. Zimmer, Das Trefferprinzip in der Biologie. *Biophysik I. Leipzig:* S. Hirzel 1947.
4. Hug, O., and A. M. Kellerer, Stochastik der Strahlenwirkung. Springer, Berlin, New York, 1966.
5. Zimmer, K. G., Studien zur quantitativen Strahlenbiologie. Akad. der Wiss. und Lit., Heidelberg, Abh. Math. Nat. Klasse (1960), Nr. 3.
   —— Studies on quantitative Radiation Biology. London: Oliver & Boyd 1961.
6. Vavilov, S.I., and T. V. Timofeeva, *J. Phys.* U.S.S.R. **7,** 9 (1943).
7. Hecht, S., S. Shlaer, and M. H. Pirenne, *J. Gen. Physiol.* **25,** 819 (1942).
8. Bouman, M. A., and H. A. Van Der Velden, *J. Opt. Soc. Am.* **37,** 908 (1947).
9. Pirenne, M. H., Quantum Physics of Vision, Theoretical Discussion. In: *Progr. in Biophysics* **2,** 193-223 (1951). London: Butler and Randall.
10. Arrhenius, S., Quantitative Laws in Biological Chemistry. London: Bells & Sons 1915.
11. Rahn, O., *Symposia Quant. Biol.* **2,** 70 (1935).
12. Jordan, P., Biologische Strahlenwirkung und Physik der Gene. *Physik.* U. **39,** 345 (1938).
    ——Die Verstärkertheorie der Organismen in ihrem gegenwärtigen Stand. *Naturwissenschaften* **26,** 537 (1938).
    —— Die Stellung der Quantenphysik zu aktuellen Problemen der Biologie. *Arch. ges. Virusforsch.* **1,** 1 (1939).
    —— Strahlenbiologie der Bakterien. *Protoplasma* **32,** 464 (1939).
    —— Zur Quanten-Biologie. *Biol. Zbl.* **59,** 1 (1939).
13. Clark, A. J., Kinetics of Drug Action. In: Handbuch d. exp. Pharmakologie, *Ergänzungsband zu Gen. Pharmacol.* **4,** 96 (1937).
14. Kellerer, A. M., Mikrodosimetrie Grundlagen einer Theorie der Strahlenqualität GSF-Bericht B1, Munich 1968.
    ——Analysis of Patterns of Energy Deposition, Proc. Microdosimetry, Stresa, Euratom (1969).
15. Rossi, H. H., Specification of radiation quality. *Radiation Research* **10,** 522-531 (1959).
16. Rossi, H. H., Energy Distribution in the Absorption of Radiation. Chapter from "Advances in Biological and Medical Physics." Vol. 11, pp. 27-85, Academic Press, 1967.
17. Rossi, H. H., Correlation of radiation quality and biological effect. *Ann. N. Y. Acad. Sci.* **114,** 4-13 (1964).
18. Smith, H. H. and H. H. Rossi, Energy Requirements and Relative Biological Effectiveness for Producing a Cytogenetic Phenomenon in Maize by Irradiating Seeds with X-rays and Monoenergetic Neutrons. *Radiation Research* **28/2,** p. 302 (1966).
19. Sparrow, A. H., A. F. Rogers and S. S. Schwemmer, Radiosensitivity Studies with Woody Plants-I Acute Gamma Irradiation Survival Data for 28 Species and Predictions for 190 Species. *Radiation Botany* 8:149-186, 1968.
20. Rossi, H. H., The Role of Associated Absorption Events in Lenticular Radiation Injury. Proc. of Second Panel on Bioph. Aspects of Radiation Quality, IAEA, Vienna, 161-170, 1968.
21. Neary, G. J., Radiation Quality and RBE for Chromosome Aberrations. Proc. of Panel on Bioph. Aspects of Radiation Quality, IAEA, Vienna, p. 58-62, 1965.
22. Humphrey, R. M. and W. K. Sinclair, The relative biological effectiveness of 22-Mevp X-rays, Cobalt-60 gamma rays, and 200-Kv X-rays. VIII. *Radiation Research* **20,** 593-599 (1963).

23. Miletic, B., Petrovic, A. Han, and L. Sasel, Restoration of viability of X-irradiated L-strain cells. *Radiation Research* **23,** 94-103 (1964).
24. Kellerer, A. M., Microdosimetry and the theory of straggling. Proc. of the Conference on Microdosimetry ISPRA, EURATOM, 1967.
25. Kellerer, A. M., Parameters of track structure. Proc. 2nd L. H. Gray Conference, Cambridge, 1969, in press. Brit. Journal of Radiology.
26. Hug, O., and A. Kellerer, Zur Interpretation der Dosiswirkungsbeziehungen in der Strahlenbiologie. *Biophys.* **1,** 20-32 (1963).
27. Till, J. E., E. A. McCulloch, and L. Siminovitch, A stochastic model of stem cell proliferation, based on the growth of spleen colonyforming cells. *Proc. Nat. Acad. Sci.* **51,** 29-36 (1964).
28. Haefner, K., *Photochem. Photobiol.* **5,** 587 (1966).
29. Kellerer, A. M. and O. Hug, Random factors in the survival curve. *Adv. in Biological and Medical Physics,* **12,** p. 353-366, 1968.
30. Paskin, A., B. V. Bronk, and G. J. Dienes, Stochastic models of cell proliferation and cell response to radiation. *Brookhaven Symposia in Biology,* **No. 20,** p. 169-178, AEC, 1968.
31. Hall, E. J. and J. S. Bedford, Dose Rate: Its effect on the survival of HeLa cells irradiated with gamma rays. *Radiation Research* **22,** 305-315 (1964).

PHYSICAL ENERGY TRANSFER AND BIOLOGICAL EFFECT

Cornelius A. Tobias*

Classically, many biological effects of radiation were thought to be associated with quantum absorption of energy. Lea and Gray (1) have emphasized the role of formation of ion pairs in biological effects. Early workers, beginning with Dessauer (2), assumed that absorption of a critical amount of energy in a definite volume could be decisive for altering biological processes at the cellular level. Difficulties arose with precise determination of the volume concept, however. Pollard (3) introduced the concept of cross section for dealing with the cellular and molecular radiobiology of accelerated, parallel beams of particles; and Zirkle and Tobias (4) showed that indirect action can also explain radiation effects in the sense that the initial, physical energy transfer may take place in a location at some distance from the crucial molecular events. Migration or diffusion of intermediate products may take place, particularly in aqueous media. Subsequently, Brustad (5) demonstrated that in dried macromolecules, indirect radiation effects still take place and can modify inactivation of enzymes.

It still remains our task to give a complete description of the manner in which initial physical interactions turn into biological events. Rossi and group (6) realized that there is a statistical distribution of energy absorption events in the kind of microscopic volumes that may be the site of biologically important events. They measured statistical distributions of energy deposition in small air-gas volumes which were dimensioned in such a manner that these volumes were mass-equivalent to the small "critical" volumes in tissue. The concept was suggested that a minimal energy absorption in the critical volume might be necessary to cause a specific effect. Oda (34,39) applied such concepts to cellular radiobiology.

Katz (7) has called attention to the fact that ionization occurs in a peculiar spatial distribution around tracks: it follows that track cross sections are not infinitely small, but should be considered as part of the overall process cross section. His method has an advantage that the medium for measurement is photographic emulsion, which in hydrated form is not too dissimilar to tissue; but a dis-

---

* The author is indebted to the Atomic Energy Commission and the National Aeronautics and Space Administration for research support.

advantage is that energy distribution can be most successfully measured some microns from the track core; whereas this quantity is of importance at much closer distances, perhaps even a few angstroms from the track core. Katz uses a smoothed radial distribution of energy deposition for modeling biological effects. The calculations of Baum (8) are of particular interest since they show that, almost independent of energy, about ½ of the energy loss of protons remains in a cylinder of 11 Å radius from the "core".

Turner and Hollister (9) in their theoretical approaches to the subject have emphasized the importance of distinguishing between primary, secondary, and higher order events. They calculated quantum transitions as the primary effect and stress the importance of correlating such results to biological events.

At Berkeley, for the past several years, we have used fast, accelerated particles for radiobiological research on molecules and cells. The techniques for using the monoenergetic beams of the Hilac have been described (10). In addition, studies were made of the range-energy relationship (11), modifying influences of straggling and multiple scattering (12).

Measurements of the fluctuation of energy deposition, the Landau-Vavilov effect, were also made (13). Biologically, it was demonstrated that the effects on dry macromolecules are exceedingly complex since they depend not only on LET but also on intra- and inter-molecular energy transfer. "Direct" effects are only demonstrable at liquid helium temperatures, at which energy deposition in a small critical volume, about $\frac{1}{10}$ of the macromolecular volume, seemed important (14). In living cells it was found expedient to use cross section determinations to account for lethality, and in some instances for mutations. The overall conclusions have been reviewed (15,16).

In this paper, correlations are made between physical energy transfer and cellular biological effect. We shall attempt to use the cell as a kind of "proportional counter" for biologically important events. We ask the question, "does the concept of depositing a minimal energy in a critical volume hold or is there some other valid representation of events?" First, however, I wish to present a sequence of events—physical, chemical, and biological—that might take place when radiation acts on living cells.

Table 1 indicates some pathways of biological effects from the point of view of cells. When survival, chromosome aberrations, or mutations are the end effects studied, the experimental evidence points strongly to the high radiovulnerability of the cell nucleus, in fact, of DNA. The effect on ribosomal and messenger RNA is present, but less important, since these molecules can be resynthesized with intact structure if intact DNA is present. Protein effects are important only as co-factors to nucleic acid damage; for example, when nucleoprotein and nucleic acids cross-link, specific protein enzymes are necessary for breakdown and repair. Thus, DNA injury, breakdown, and subsequent rejoining and repair are salient features of the action of radiations on cells; and we shall focus on these. One should mention that radiation also has profound effects on cell membranes and that, on special systems, injury to these can be a determinant to subsequent biological events. For example, it is believed that interphase death of lymphocytes and of thymocytes following

TABLE 1.

PHYSICAL INTERACTION—ATOMIC LESION
- Direct
- Indirect
- Ionization
- Excitation

PRIMARY MOLECULAR LESION
- Free radicals
- Excited molecules
- Dissociation

MODIFIED MOLECULAR LESION
- Hydrogen peroxide
- Organic peroxides
- Chemical modifiers
- Activated enzymes

DNA LESIONS
- Single chain scission
- Protein-DNA link
- Base alteration
- Base-to-base link
- Double chain scission
- Chain crosslink

CELLULAR EVENTS
- Enzymatic breakdown of DNA
- Enzymatic repair of DNA
- Cell division process
- Faulty transcription
- Faulty translation
- Faulty polymerization
- Faulty self-duplication

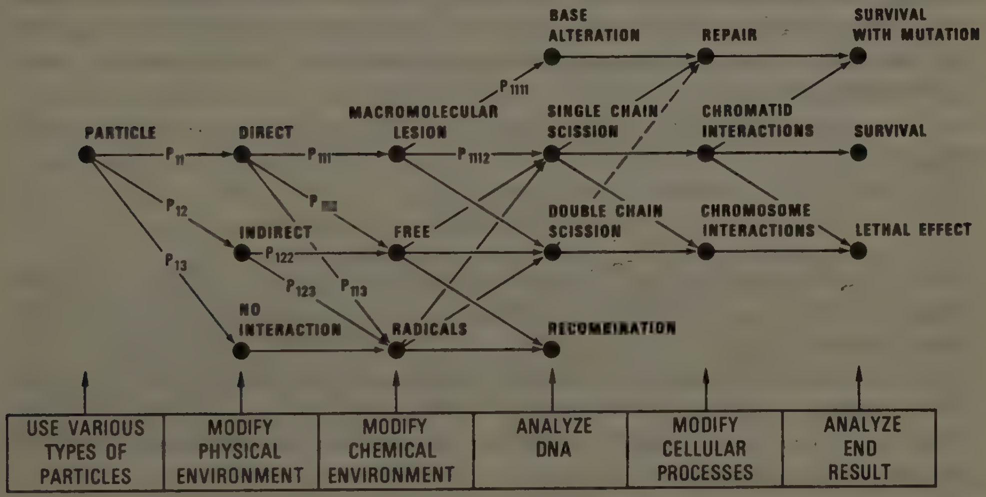
GROSS OBSERVATIONS
- Lethality
- Cell division inhibition or delay
- Mutation

radiation injury may be due to primary injury and subsequent enzymatic breakdown of membranes (17).

It is our prime interest to establish how nucleic acids can be injured *in vivo* or *in vitro,* what the various types of lesions are, which lesions lead to specific biological effects, and which are modified by repair processes. DNA lesions of sufficient stability for experimental study have been identified as base alteration, base-to-base interaction, single chain scission, double chain scission, cross-link and protein-DNA link. It is sometimes assumed that all of these effects, except double chain scission, appear to be repairable by various degrees though this has not been shown with reliability for each type of lesion in a wide variety of cells.

A model for successive stages in developing lesions is shown in Figure 1. This figure also gives clues with respect to the action of chemical and physical modifiers of radiation effects. For quantitative handling of radiation data, it may be worthwhile to note at this point that the number of individual molecules (or cells) that are in later stages of radiation injury always depends on the number that have reached an earlier stage. In the language of probability theory, we can ask the question, "what are the alternative possibilities for a lesioned cell in state X?" The states of the cell between an initial physical event and the appearance of gross injury can usually be described as a discrete branching Markov process.

This is a simplification. There seems to be a relatively small number of different types of molecular lesions and finite number of alternatives for each successive state. However, there appears to be continuous change in parameters of radiation sensitivity in different stages of the cell division cycle. We shall attempt to simplify the description of the process by averaging parameters for many physiological states. If we know the initial and final stages of interest, the process of arriving from an initial tc a final state may be characterized by a product of probabilities if a single pathway is available and by a sum of products if alternative pathways exist.


BASE ALTERATION
REPAIR
SURVIVAL WITH MUTATION
P_{1111}
MACROMOLECULAR LESION
SINGLE CHAIN SCISSION
CHROMATID INTERACTIONS
PARTICLE
DIRECT
SURVIVAL
P_{11}
P_{111}
P_{1112}
P_{12}
P
DOUBLE CHAIN SCISSION
CHROMOSOME INTERACTIONS
INDIRECT
FREE
LETHAL EFFECT
P_{122}
P_{13}
P_{123}
P_{113}
NO INTERACTION
RADICALS
RECOMBINATION
USE VARIOUS TYPES OF PARTICLES
MODIFY PHYSICAL ENVIRONMENT
MODIFY CHEMICAL ENVIRONMENT
ANALYZE DNA
MODIFY CELLULAR PROCESSES
ANALYZE END RESULT

DBL 708 5873

Fig. 1. Discrete Markov chain model of some of the events that may occur as a result of the passage of a particle through a living cell. Usually it is feasible to measure the end result in terms of probabilities for survival, lethality and mutation. These probabilities are actually appropriate products and sums of the branching probabilities. It has been possible to modify the process probabilities at several levels as shown on the Figure. The actual situation is obviously much more complicated than what is shown here. When detailed pathways are not fully known, it is possible to lump process probabilities, between any two points of the graph, into a single value.

One must remark here that the values of each branch probability depend on the physio-chemical state of the environment. There is probably not a single case of biological effect known for which complete knowledge exists for all the necessary branch probabilities. However, experiments with different environments allow us to alter the branch probabilities; and, in some instances, to evaluate their magnitude.

We shall now take some examples from a series of experiments carried out at Berkeley on the cellular effects of accelerated heavy ions and attempt to give a semi-quantitative analysis of the steps involved, hopefully arriving back to identify some of the physical interactions that may have caused the events.

INHIBITION OF COLONY FORMATION

It has been amply demonstrated that survival of colony formation in microorganisms irradiated when they are isolated from each other is either an exponential function of the flux density of particles, or is a function made up of exponential probability terms commensurate with Poisson's theory of random events and discrete chains of Markov processes. A glance at Figure 1 will convince us, however, that many conditions (and therefore several probabilities) may be altered by performing the experiment in different physical, chemical and biological conditions. We would expect that the survival equation would have coefficients that express the change in survival in response to one of the changed environmental or radiation variables.

A simple exponential survival curve may be characterized as function of dose, *D*, as

$$N = N_0 e^{-s.D}, \tag{1}$$

where N is the number of survivors from an initial number, N_0, and s is the inactivation constant. For the statistical treatment we wish to present, it is more suitable to find another variable instead of dose, D, that is independent and may be regarded as completely random. Dose is not an independent variable since it depends on flux density, f, of the parallel particles impinging on the cells and on the energy transfer, ε, from each particle[1]. Since we always have a distribution of energy transfers, this must be taken into account. The dose is

$$D = \int_{\varepsilon_{\min}}^{\varepsilon_{\max}} \varepsilon . f(\varepsilon) d\varepsilon, \tag{2}$$

---

[1] The use of energy transfer $\varepsilon = \frac{dE}{dx}$ has also been criticized, since there are different particles with equal energy transfer but different atomic numbers and velocities. These have different distribution of ionzation energy around the track core. There is also an uncertainty of whether or not one should use the entire physical energy transfer or the energy transfer to some small radius, r_0, around the track core. In the treatment given here either can be used. They will merely give rise to different apparent cross sections.

and the quantity $s.D$ may be expressed with the use of cross section $\sigma(\varepsilon)$,

$$s.D = \int_{\varepsilon_{\min}}^{\varepsilon_{\max}} \sigma(\varepsilon).f(\varepsilon)d\varepsilon \tag{3}$$

For the case when we have strictly monoenergetic particles and assume that each has the same LET, ε, and therefore neglect statistical fluctuations in ε, we have

$$sD = \sigma(\varepsilon).f^*(\varepsilon) \text{ and} \begin{cases} f^*(\varepsilon) = D/\varepsilon, \\ \sigma = s.\varepsilon \end{cases}$$

where $f^*(\varepsilon)$ is the integral flux density and $\sigma(\varepsilon)$ is the overall product of the probabilities in the branches of the Markov chain that contribute to the effect. This includes the cross section of the part of the biological specimen that must be crossed to cause lethal effect and is modified by probabilities for energy deposition as the object is crossed as well as probabilities for repair. Sometimes, when more than one molecular lesion is involved, several branch probabilities are of importance.

A number of macromolecules and cellular systems have been studied by various investigators at our laboratory and elsewhere with the track-segment method. These include enzyme molecules (5,14), phage particles (18,19), bacterial spores (20), yeast cells (21), mammalian cells (22,23), *Micrococcus radiodurans* (24), and plant seeds (25). Although each such system is very complex and has special properties, one can make some generalizations that seem to apply to several. The special cases can then be studied and an attempt made to explain special behavior where it differs from the norm. The reader is referred to individual publications for full details and to previous attempts to systematize the results (15,16).

For most of the above experiments, whether radiation exposure was dry or wet, at low temperatures or at room temperature, in the presence of air or in nitrogen, a general expression was found to express the LET-dependent cross section. This expression fits the experiments on spores of *B. megaterium* best where the most accurate data are available,

$$\sigma = \sigma_A.P_A(\varepsilon) + \sigma_B.P_B(\varepsilon) \tag{5}$$

Here σ_A and σ_B are constants having dimensions of geometrical cross section, whereas P_A and P_B are probability factors. For lethal effects, P_A and P_B approach unity when ε becomes very large, so σ_A and σ_B are limiting cross sections for very large ε. Writing out the empirically found expression for *B. megaterium* in more detail, we have

$$\sigma = \sigma_A(1-e^{-\alpha\varepsilon}) + \sigma_B(1-e^{-\beta^2\varepsilon^2}) \tag{6}$$

In the above, α,β are constants at constant environment, independent of LET. Environmental modifications during or after exposure (for example, temperature, gas environment, moisture content, and presence of alcohols and of other chemical and biological modifiers) do modify these constants. We regard the first expression as due to a "single ionization event" and the second term as due to "co-operative interactions". That the two terms represent, indeed, different processes may be argued on their radically different dependence on LET: The first term depends on ε and the second on ε^2. In addition, evidence is accumulating on qualitative differences between the effects represented by the two terms.

The purpose of this paper is to inquire about the relationships between physical energy transfer and biological events. We now try to illuminate this problem by two different methods: by using Poisson probability distributions to identify the natures of the two different types of events, and by an attempt to relate these to the statistical distributions of energy transfers along the track.

PROBABILITY APPROACH TO THE "SINGLE EVENT" PROCESS (TYPE A)

Assume that production of a single primary ionization (that is, production of a single delta ray) along the straight line trajectory of the track of the bombarding ion falling within a critical distance, t, along the track is sufficient to cause a biological event of type A. If the average energy loss per delta ray is E_δ then the average number of delta rays per cm is

$$n_\delta = \frac{\varepsilon}{E_\delta}$$

When the particles cross a slab of matter of thickness, t, perpendicularly to the slab, the average number of primary events is

$$\text{Av.no.} \approx t.n_\delta = \frac{t\varepsilon}{E_\delta} = \alpha\varepsilon \qquad (7)$$

Here $\alpha = \frac{t}{E_\delta}$ is a constant. When the Poisson distribution for random events is used the probability that one or more delta rays are found in thickness t is

$$P_A = 1-e^{-\alpha\varepsilon} = 1-e^{-\frac{t\varepsilon}{E_\delta}} \qquad (8)$$

This expression then gives the efficiency of primary ionization in thickness t along the track. The overall cross section will be, as in the first term of equation (6),

$$\sigma(\varepsilon) = \sigma_A(1-e^{-\alpha\varepsilon})$$

The use of accepted values of E_δ (e.g., about 100 ev per delta ray) allows estimation of the "critical thickness," t from the empirically determined constant α. In this model, our requirement is to have at least one ionization event in thickness t, in an area determined by combined geometrical factors of the cell nucleus and the particles.

RELATIONSHIP OF THE SINGLE HIT EVENTS TO STATISTICAL VARIATIONS OF ENERGY TRANSFER

Since energy transfer events are discrete ionizations or excitations, it is known that in a thin slab of matter energy deposition undergoes statistical fluctuations. As a consequence, the energy transferred from each particle may be different. Landau and Vavilov have made quantitative models for the statistical fluctuations in energy loss, and some measurements have been carried out recently in this field. In our laboratory, energy loss distributions from heavy accelerated ions were measured in solids (13), whereas at Columbia University (26) such studies were carried out in air with a proportional counter. It is difficult to measure distributions of this type in slab thicknesses that are biologically meaningful since the critical thickness in biological material in some cases extends to only a few angstroms. However, we know that the frequency distribution of energy transfers through a thin slab of matter has the characteristic shape shown in Figure 2. Assume that the frequency of energy transfers of magnitude E in a thin slab of tissue of thickness, *t*, is *L* (*E*,*t*). The sum of frequencies is unity:

$$\int_0^\infty L(E,t)\,dE = 1 \qquad (9)$$

The mean energy transferred in slab t is $\varepsilon.t$. Otherwise this is given in the expression

$$\varepsilon.t = \int_0^\infty E.L(E,t)\,d\mathrm{E} \qquad (10)$$

Assume now that a biological effect is observed due to a single particle only, if the particle crosses the area σ_A, and if it transferred more energy to the slab of tissue than a minimum, E^*, necessary to produce the effect. Energy transfers which are less than E^* are assumed to cause no effect. The probability, P_A, that a particle confined within σ_A will do this is

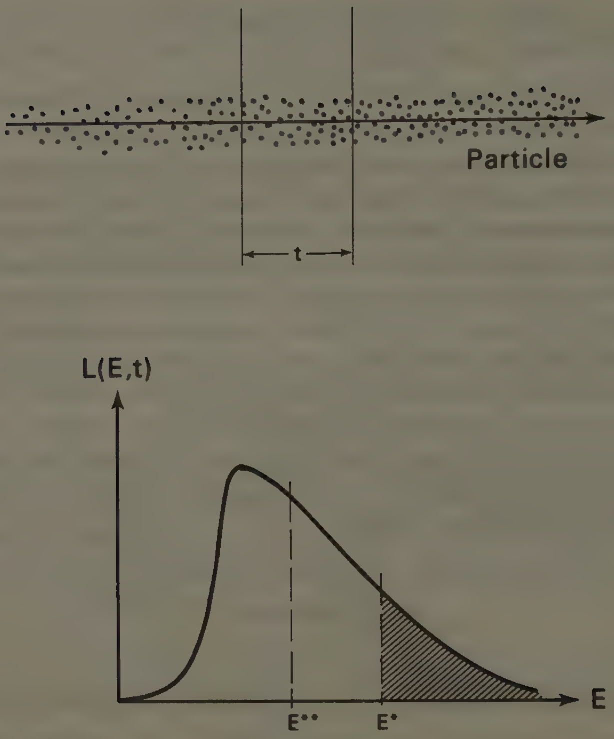
$$P_A = \int_{E^*}^\infty L(E,t)\,d\mathrm{E} \qquad (11)$$

For single hit events, empirically we found from the first term of (6)

$$P_A = 1 - e^{-\varepsilon\alpha} \qquad (12)$$

We shall use this expression to obtain an analytical form for $L(E)$. Substitute for the value of the constant alpha, which has dimensions of $\frac{1}{\epsilon}$, $\alpha = \frac{t}{E^*}\cdot$

Next, differentiate P in equations (11) and (12). We obtain:



DBL 708 5872

Fig. 2. This drawing represents the passage of a particle through a slab of matter of thickness t, representing a biological target, e.g., a chromosome or a DNA nucleoprotein molecule. The energy, E, transferred to tissue shows statistical variations characterized by the distribution L(E,t). The typical form of distribution is shown in the lower part of the drawing. The calculation assumes that energy transfer in excess of E* causes a biological effect, whereas if it is less than E*, the organism would survive. E** might be another energy limit for the production of mutations. A given type of mutation might occur if the energy transfer is between E** and E*.

$$L(E^*,t) = \frac{-d}{dE^*} \int_{E^*}^{\infty} L(E,t)\,d\mathrm{E} = \frac{t\varepsilon}{E^{*2}}\, e^{-\frac{te}{E^*}} \tag{13}$$

The general form of this expression $L(E)$ is in near agreement with most of the physical theoretical approaches to the problem of energy transfer. It has a nonlinear rise at low E and falls off as $1/E^2$ at high E. There are two adjustable constants, t and E^*. Only one of these can be determined from the experiment, usually the ratio $\frac{t}{E^*} = \alpha$.

In Figure 3 several *L* distributions are drawn. Some of the empirically determined values for alpha are given in Table 2. If we assume for E^* the average energy required in air for producing a single delta ray, we obtain for t the values listed in the last column of the table.

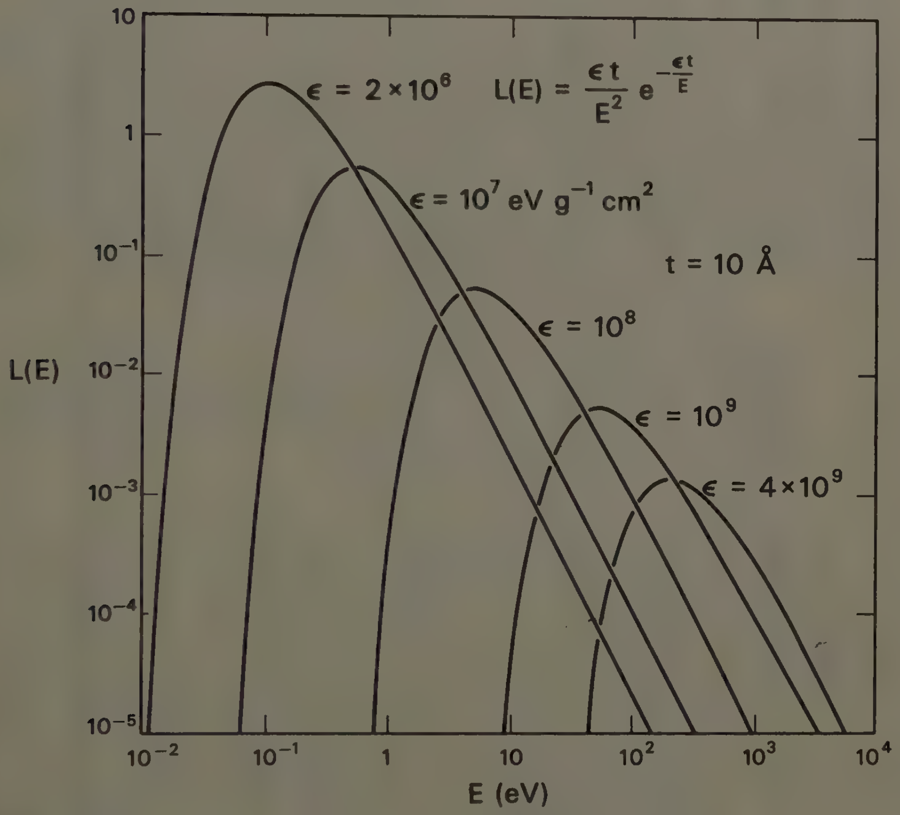


Fig. 3. Calculated values for normalized L (E) distributions. Each curve is for a different value of $t\epsilon$.

RADIATION INACTIVATION OF ENZYMES

The inactivation of dried polycrystalline enzymes has been particularly carefully studied by Brustad. At a given temperature cross section curves may be obtained that fit expression (6) reasonably well, but at different temperatures the constants of the expression change. We believe that the explanation for this is that transfer of excitation energy can occur at all but the lowest temperatures, near that of liquid He. The cross section increases as the irradiation temperature is increased. We believe that below liquid nitrogen temperature, intra-molecular energy

Table 2. *Cross Sections and Constants in Function $P(\varepsilon)$, $L(E)$ and $K(E)$ as Determined on the Basis of Experimental Data**

Organism	Single Hit Effect			Cooperative Effect		
	$\sigma_A\ cm^2$	$a[cm^{-1}Mev]$	t_α *Angstrom if* $E^* = 100ev$	$\sigma_B\ cm^2$	$\beta[cm^{-1}Mev]$	t_β *Angstrom if* $E^* = 100ev$
Dry B. Megaterium spores in air[20]	4.5×10^{-11}	6.35×10^{-3}	63 Å	1.79×10^{-9}	6.7×10^{-4}	6.7 Å
Dry B. Megaterium spores in nitrogen[20]	4.6×10^{-11}	4.6×10^{-3}	46 Å	1.62×10^{-9}	6.0×10^{-4}	6.0 Å
Wet haploid yeast cells in air[31]	1.2×10^{-9}	2.6×10^{-3}	28 Å	1.08×10^{-8}	6.9×10^{-4}	6.9 Å
Wet haploid yeast cells in nitrogen[31]	10^{-9}	1.15×10^{-4}	12 Å	0.9×10^{-8}	6.4×10^{-4}	6.4 Å
Wet kidney T_1 cells in air. Exponential component[22]	9×10^{-8}	4×10^{-4}	4 Å	81×10^{-8}	4.2×10^{-4}	4.2 Å
Wet kidney T_1 cells in air. Components capable of recuperation[22]	14×10^{-8}	3×10^{-4}	3 Å	14×10^{-8}	6×10^{-4}	6.0 Å

*These values supercede our earlier quoted numbers. They were either taken from the references given or recalculated for the purpose of this report.

transfer occurs; and above liquid nitrogen temperature, inter-molecular energy transfer may modify inactivation.

Most of the radiation effects on phage particles appear to be explainable by this mechanism even at high LET, and most of the effects on cells at low LET are also accounted for by it. There is some evidence that the ionization events are followed by formation of free radicals, H· and OH·, and that these are responsible for most of the effects in aqueous phase at low LET. The action of oxygen in modifying these effects is fairly well understood.

The effects of chemical modifiers, like oxygen, and of sulfihydril substances appear to depend directly on the constant α. Energy transferred in nitrogen is not used completely efficiently for producing molecular lesions: in the presence of oxygen and by interaction of oxygen with free radicals present, more of the energy transferred becomes effective. By studying the coefficient t/E^* at different oxygen concentrations, one may obtain a concentration dependence of the changes. These effects also depend on pH, since the interaction of oxygen with free radicals depends on pH. The simplest expression that accounts for the changes in the presence of oxygen is to replace t/E^* with the expression

$$\alpha(O_2) = \frac{t(N_2)}{E^*(N_2)} \cdot \frac{1 + a[O_2]}{1 + b[O_2]} \qquad (14)$$

when irradiation occurs in pure nitrogen,

$$\alpha(N_2) = \frac{t(N_2)}{E^*(N_2)},$$

and when the oxygen concentration $[O_2]$ is high enough, usually above 1% in the gas mixture, then $\alpha[O_2] = \frac{t(N_2)}{E^*(N_2)} \cdot \frac{a}{b}$; a and b are constants of the oxygen interaction. Unfortunately, we do not know whether only E^* changes as function of the oxygen present, or if it is necessary to change t also. Since the presence of oxygen changes the composition of free radicals, causing the appearance of $HO_2\cdot$ instead of some H•, it may also alter the recombination rate of H• and OH• and change the mean free path and lifetime of the radicals present. Thus, both minimal energy deposition and the critical distance may change. Unfortunately, there are not enough precision studies available on the effect of the oxygen as a function of concentration. Type A effects are in large part annealable if irradiation is carried out in dry spores and in nitrogen. Subsequent treatment while still in the dry state with heat or with post-irradiaton H_2S appears to remove almost all of the effects of this type, as evidenced by the large decrease in the cross section shown by Powers *et al.* (20). At least 80% of the effect can be annealed in this manner, and this process appears to be similar to annealing of induced free radicals produced by radiation in crystalline solids. After water is admitted to the system, heat or H_2S annealing is no longer effective.

We may surmise that at least part of the type A process is eventually expressed as single strand breaks in DNA and is subject to post-irradiation enzymatic repair. Such repair processes would modify the cross section. Thus, an overall expression for the cross section is

$$\sigma = \sigma_A \,.\, K_{\text{anneal}} \,.\, K_{\text{repair}}, \left[1 - \exp\left(-\frac{\varepsilon . t}{E^*} \cdot \frac{1 + a[O_2]}{1 + b[O_2]}\right)\right] \qquad (15)$$

The expression in brackets is the probability of a molecular lesion in path t. It is assumed here that the amount of annealing and of repair are proportional to the number of lesioned molecules. The annealing and repair constants, K_{anneal} and K_{repair}, depend on a number of factors; e.g., temperature, time and physiological conditions.

COOPERATIVE EVENTS

The lethal effects caused by high LET radiation have a cross section dependence which follows the second part of Equation (5) or some similar expression. Ninety percent of the lethal effects at high LET on living cells with relatively large chromosomes appears to be due to this mechanism.

The cooperative nature of the interaction is evident from the simple statistical argument which leads to the derivation of formula (6). Assume that along a single track in a location essential to the integrity of the cell at least two separate sets of ionizing events occur, perhaps in two different locations; and that some of these events interact in a very short time interval to produce a new species of product K. The average number of K molecules formed should be proportional to the number of primary ionization events in each location. The number of primary ionization events is in turn proportional to the average rate of energy transfer, ε; namely, $a_1\varepsilon$ and $a_2\varepsilon$. Thus,

$$N_K \approx a_1\varepsilon a_2\varepsilon \approx \beta^2\varepsilon^2 \text{ where } \beta^2 = a_1 a_2 \qquad (16)$$

Using the Poisson distribution, the probability that at *least* one product molecule should be formed is:

$$P_B = 1 - e^{-\beta^2 \varepsilon^2} \text{ and } \sigma = \sigma_B P_B \qquad (17)$$

which is identical with the second part of Equation (6) and reasonably closely fits the empirical observations on bacterial spores and mammalian cells.

DISTRIBUTION OF ENERGY TRANSFERS

We may proceed in a manner similar to the above with the type A mechanism and ask if there is a significant distribution of physical events in a single track segment that lead to the probability expression (17). Assume that the frequency of such events is $K(E,t)$ and that the distribution is normalized so that

$$\int_0^\infty K(E,t)\,dE = 1 \qquad (18)$$

Is there a critical energy E^* such that all events above this critical energy in track segment t cause a cooperative event K? This is true if

$$P_B = \int_{E^*}^{\infty} K(E,t)dE = 1 - exp\,(- t^2\varepsilon^2/E^{*2}). \tag{19}$$

Here we assume that $\beta^2 = \frac{t^2}{E^{*2}}$ and that E^* is the critical energy. We obtain the analytical form of K by differentiating.

$$-\frac{dP_B}{dE^*} = K(E^*,t) = \frac{2t^2\varepsilon^2}{E^{*3}}\, exp\,(- t^2\varepsilon^2/E^{*2}). \tag{20}$$

This expression, which depends on ε^2 explains why the cross section is increasing so rapidly with increased LET. The integral

$$-P_B(\varepsilon) = \int_{E}^{\infty} \frac{2t^2\varepsilon^2}{E^3}\, exp\,(- \beta^2\varepsilon^2/E^2)\,dE \tag{21}$$

increases rapidly with increasing LET; first as the square of LET and later leveling off to a plateau of 1 at high LET.

The K distribution does not correspond to primary ionizations which seem to follow the L distribution more closely. In fact, a physical product of energy transfer that has such a distribution is not known. It is likely, however, that H_2O_2 formation along high LET tracks in water does follow a similar relationship to Eq. 21. In any case, if one wished to obtain direct physiochemical evidence for the cooperative phenomenon one should look for it as having a yield along tracks specified by Eq. 21. There is some evidence that free radical spectra observed in dry protein (27) and in dried bacterial spores show a new component at high LET that is qualitatively different from the component observed at low LET and that the yield of this free radical component increases at high LET. However, the problem of quantitation of these free radicals is difficult, particularly since at high dose levels, radical formation is accompanied by radical destruction (28).

Calculated K(E) distributions are shown in fig. 4.

The empirically determined coefficients for lethal effects on cells are given in Table 2. The data allow determination of $\beta = \frac{t}{E^*}$ only. It is interesting that beta values for different cell strains do not differ from each other markedly. LET dependence of the cooperative effect is similar, and sharp rises occur at about the same LET for three different cell strains (yeast, *B. megaterium,* and mammalian cells). An explanation is that a similar molecular structure is involved in radiation action in each strain of cells. If we assume a value for E^* of about 100 ev, we obtain about 10 angstroms for t. Although assignment of such dimensions has chiefly speculative value, this value of t corresponds nearly to the diameter of the DNA double helix. Thus we may visualize that for cooperative lethal action, the same particle track needs to simultaneously deposit two delta rays, generated less than 10 angstroms from each other and interacting in such a way that both strands of DNA could be simultaneously affected. This is in agreement with a prevalent idea

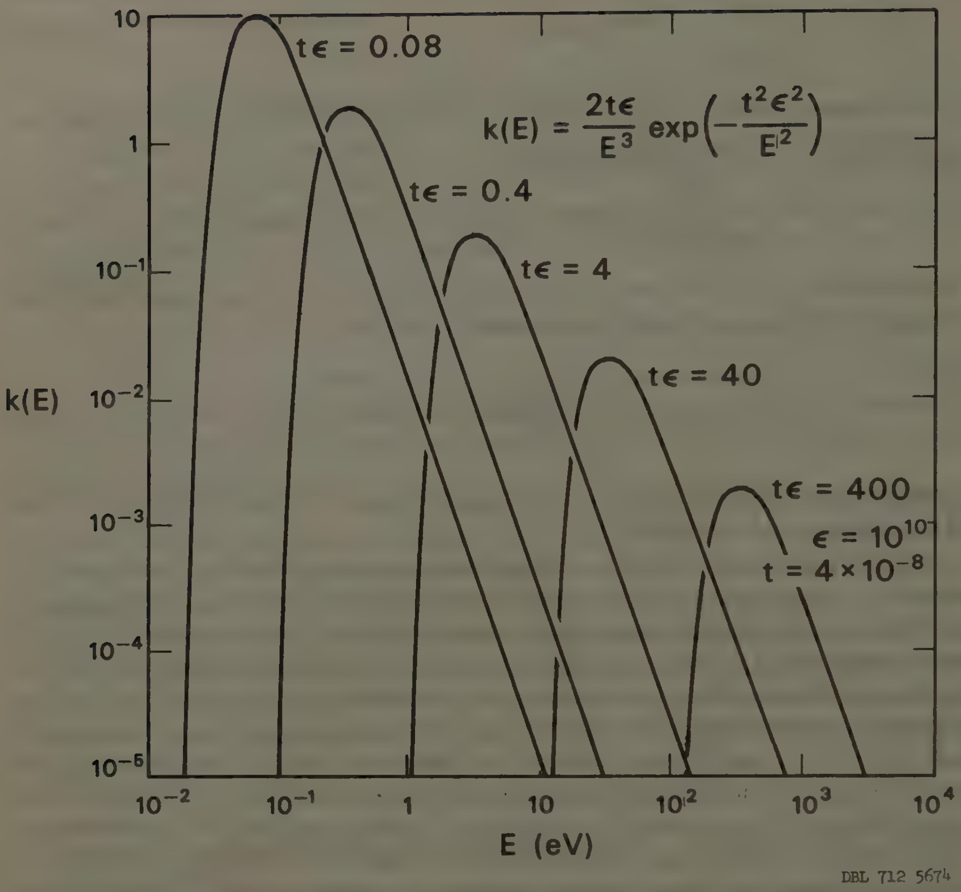


Fig. 4. Calculated values for normalized K (E) distributions. Each curve is for a different value of $t\epsilon$.

that high LET particles exert their lethal effects primarily by producing double strand scission in DNA. Dewey, working in our laboratory with argon ions and bacteriophage, has shown that the part of lethal damage that is not modifiable by SH group protectors corresponds to double strand scission. More work is being done on this problem at present, and it will be necessary to extend the techniques of strand scission measurements so that they can be applied to cells of highly differentiated tissues, including mammalian and plant cells. The work of Cole and of Elkind is showing encouraging progress in this direction (29,30). If high LET radiation does, indeed, exert its lethal effects by the production of double strand breaks, this would help explain its high biological effectiveness. It would take many low LET particles to produce two individual, single strand breaks, in juxtaposition to each other, to break both strands. Whereas a single, high LET particle can cause killing, it takes hundreds or sometimes thousands of low LET particles passing through the same cross section to cause lethality.

COEXISTING EFFECTS

On occasion we have used another model to explain some of the high LET effects by assuming the production of two or more radiation lesions, along the same ionizing track, which are formed independent of each other. Unlike the case for cooperative effects where it was required that ionization products from two parts of the same track should interact in a very short time interval, here we require merely that they should be simultaneously present in segments of the track.

Let the average number of decisive events along a track be $\beta\varepsilon$. Using the Poisson distribution we shall now ask what the probability is that two such events be formed simultaneously along the same track in segment t. This is

$$1 - (1 + \beta\varepsilon)e^{-\beta\varepsilon} \tag{22}$$

By the same kind of reasoning we can form an energy distribution K_x for such paired events and obtain

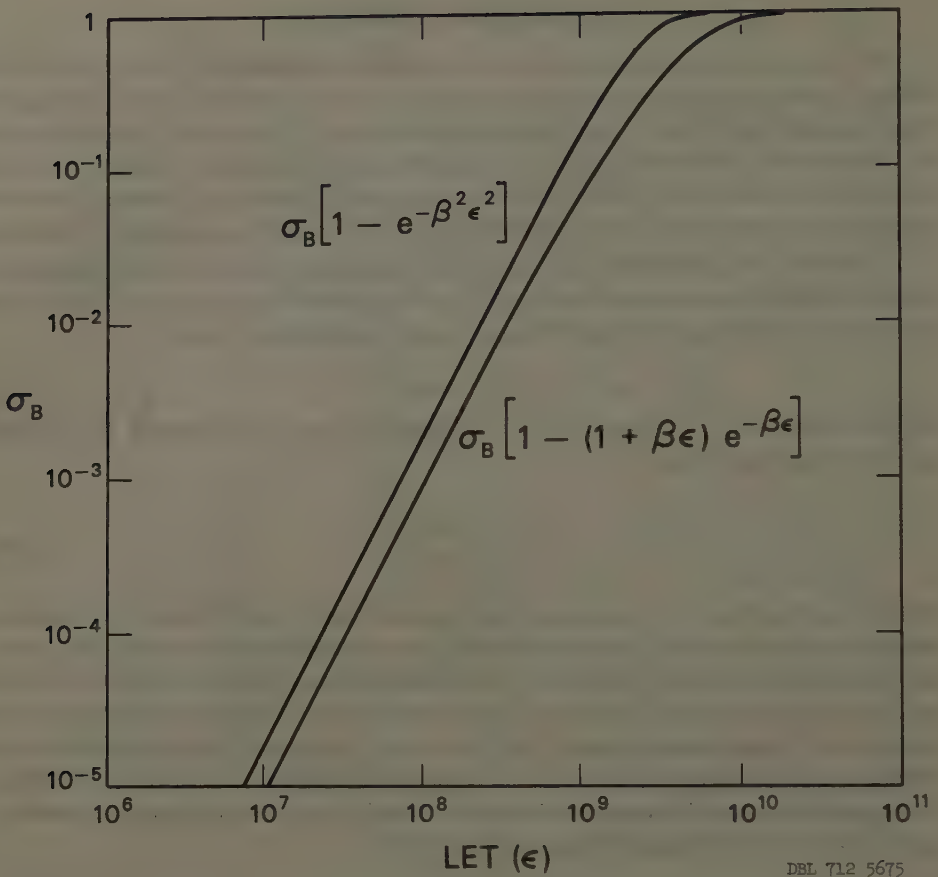
$$K_X = \frac{t^2\varepsilon^2}{E^3} e^{-\frac{t\varepsilon}{E}} \tag{23}$$

The K distribution for cooperative effects and the K_x distribution for coexisting effects are rather similar: at low LET they cause the cross section to rise proportionally to E^2. At high LET, however, the K_x distribution reaches plateau slower than the K distribution (see Figure 5). We find that *B. megaterium* data seem to follow the K distribution better,while the events related to tumor induction in arabidopsis, as studied by Smith *et al.*, (25) appear to follow K_x. In order to be able to distinguish between the validity of the two distributions it is necessary to measure, with precision, cross section values at high LET.

INDUCTION OF MUTATIONS

Heavy ions are very efficient in causing chromosome breaks and rejoinings; many of these presumably involve double strand scissions in DNA.

If we exclude mutations that are due to chromosome deletion or gross chromosome rearrangements and deal with "point" mutations which correspond to small changes of coding at the level of one or two base pairs, it is generally assumed that the production of such mutations may involve smaller energy transfer than the production of lethal effects. As the LET increases, the yield of such mutations usually decreases since the high LET particles are becoming increasingly efficient in producing chromosome breaks or lethality. It seemed of interest to see if the "single event" hypothesis is sufficient to predict such effects or if we need the cooperative K or coexisting K_x models. Referring to Figure 2 we may assume that a mutation is produced if the energy transfer in track segment t is smaller than that necessary for the production of lethal effects, E^*, but larger than another critical energy, E^{**}. Using the L distribution we may calculate the LET dependence of the σ_M cross section of such mutations as



DBL 712 5675

Fig. 5. LET dependent cross section values for cooperative effects (K distribution) and coexisting effects (K_x distribution). With the same constant β, the behavior of the curves is markedly different at high LET. Precise measurements are needed at high LET to determine which of the distributions apply.

$$\sigma_M P_M = \sigma_M \int_{E^{**}}^{E^{*}} L(E,t)\,dE = \sigma_M\,[exp\,(-t\epsilon/E^{**}) - exp\,(t\epsilon/E^{*})]\,. \qquad (24)$$

Heavy ions produce such mutations less efficiently than light ions. From the nature of the single event hypothesis, we may assume that oxygen and chemical modifiers are effective and that repair can also occur. Physically, one may picture such lesions as being produced in the track penumbra rather than at the track core, perhaps due to the delta rays. The RBE of such lesions is always less than 1. Actually, when point mutations are experimentally tested, one often finds an increase in RBE, followed by a rapid decline as shown by Mortimer *et al.*, (21) in yeast. To help explain this, we should invoke the K or K_x mechanisms discussed above. The K_x mechanism should be invoked in cases where the RBE rises above 1, but the mutation induction remains oxygen sensitive even to very high LET. In a generalized case of this kind, one may assume that one, two, three, or more events in segment t of the ion track could each lead to lethality or mutation, but that the probability of mutation induction is decreasing as the multiplicity of events increases with increasing LET; whereas the probability for lethal event or complete scission of DNA is increasing. A cross section formula for mutations that fits these ideas as well as the observed data is:

$$\sigma = \sigma_M\left(a_1\varepsilon + \frac{a_2\epsilon^2}{2!} + \ldots\right)e^{-\alpha\varepsilon} \qquad (25)$$

The constants a_1, a_2 and σ_M must be determined by empirically fitting data on mutation induction.

DISCUSSION

We have atttempted to demonstrate that experimentally observed cross section data, observed with different LET radiations, can be used to demonstrate some of the statistical properties of the physical energy distribution of ionizing events that relate directly to the biological effect observed. Thus, living cells can augment the information obtained with proportional counters to give us indications of the type of primary physical events involved. It has been demonstrated that there are a number of steps between primary physical events and eventual biological observations, so that inferences of the type demonstrated here must be handled cautiously and will never replace basic physical and chemical investigations, which must be carried out with precision before we can completely understand the mechanisms of radiation effects.

Most of the low and high LET effects on cell survival and colony forming ability are explainable by assuming that two kinds of processes are operating: In the single event hypothesis, it is assumed that a single delta ray or possibly an ion pair is able to cause the effects. This process can probably do no more than cause directly, or indirectly, a single chain scission in DNA, which can be modified by chemical modifiers, including oxygen, and is subject to enzymatic repair. It is probable, from work with phage, that enough such events must accumulate to produce a double chain break before significant inactivation can occur; 90% or more of the effects of ionizing radiations on enzyme molecules and bacteriophage seems to be due to such a process, even at high LET. The process itself is in most cases (except, perhaps, at liquid helium temperatures) indirect in the sense that the initial physical energy transfer does not need to be at the molecular location where final irreversible damage occurs. Indeed there is some evidence that—in the dried state—excitation energy transfer, and—in aqueous systems—free radical production and migration play an essential role.

Since the single hit type of event can be a single ion pair, the delta rays of heavy ion tracks fully participate and the presence of a track core has no special significance except in the sense that energy is wasted in it.

Inactivation more characteristic of the track core occurs when cooperative or coexistent models are tested. Cooperative effects appear to be responsible for most of the biological effects of high LET particles and their high biological effectiveness. They rise proportionally to the square of the LET before they plateau. The simple idea that an energy deposition greater than a critical E^* in a micro-volume might cause such effects does not appear to fit. Instead, cooperative effects depend on the interaction of at least two independent ionization events.

It is tempting to assume that the end result of cooperative effects is double chain scission of DNA; however, this is not completely proven. We do know that usual chemical modifiers have relatively little effect, and it appears that usually little if any repair of such lesion occurs. It is likely that chromosome deletions and rearrangements are very efficiently caused by the cooperative effect.

Cross sections for lethality by the cooperative effects seem to be of the order of magnitude of the cross sections of cell nuclei; but they are, perhaps, somewhat larger. Since, at low dose rates the majority of effects due to single ionization events are known to be repaired, it is possible that most of the lasting chromosome damage at low or high LET is due to the cooperative effect. There is some evidence that the cooperative effect may also involve indirect action. In haploid yeast, glycerol seems to protect against it (31).

Calculation of the Effects of Mixed Radiations

Expressions 5 and 6, as well as the L and K distributions, can be used with some confidence for describing the effects of monoenergetic particles with a discrete specified rate of enrgy loss ε.

A more detailed consideration indicates that a dose of mixed radiation field with a distribution of ε values does not exert the same biological effects as the same dose of a radiation at a "mean LET" arrived at by arithmetic averaging. Rather, a mixed radiation field can be more nearly represented as a characteristic mixture of effects of type A with LET ε_1 and of type B with LET ε_2. (To be discussed in detail elsewhere.)

Search for Delta Ray Effects

It would be important to know what portion of the biological effects of heavily ionizing radiations is due to the "track core" and how much of the effect is due to the penumbra, usually called the delta ray effect. This problem can be studied by exposing the same biological specimen at the same LET to particles of different atomic number, moving at different velocities. The experience so far, using particles up to 10 MeV/nucleon, has been largely negative; and the role of delta ray effects in cell lethality has not been clearly settled. There are some differences in cross section values for hamster kidney cells obtained at high LET studied by Barendson (23) and Todd (22). The former author used low energy helium ions for some of the LET exposures that were obtained with higher energy carbon and neon ions by the latter author. The comparisons in this work of the effects of oxygen also yielded different values. The higher velocity heavier ions have more of their ionization deposited in the track penumbra and appear to exhibit greater oxygen effect. Stanley Curtis (32) has attempted to explain the differences by introducing a velocity dependent factor related to track structure for the explanation of the oxygen effect.

At very high LET, the track core contains quite high energy density. In Todd's work with kidney cells, an indication has been obtained (22) that the models here described do not fit and that the cross section for lethality is higher than expected. This is not completely surprising since various authors suggested that very heavy ionization has special effects. A. Norman (33) discussed the properties of the "thermal spike" model, though other investigators doubt the validity of this approach (28). In mica and in plastics, very heavy ions cause permanent damage of a kind that does not seem to occur at low rates of energy loss. Such effects can be made visible by post-radiation treatment of the materials with sodium hydroxide (35,36).

If radiation effects are due to the track penumbra ("delta rays") one expects them to exhibit marked oxygen effect even at very high LET along with a declining RBE and cross section. In our analysis it appears that certain yeast mutations fulfill these criteria. Here, however the "coexistent" effect of more than one ionization event must be involved to explain a peak in the RBE. Thus it appears that the track core of a very high LET particle is likely to cause lethality via the cooperative effect, while the track penumbra can be efficient in the production of point mutations in survivors.

Some Anomalies

At low LET, often several molecular lesions must be caused by several particles for lethal effect, as is the case in survival for several strains of mammalian cells. Often a single lesion of the cooperative type is sufficient to cause lethality, so, at high LET, multihit survival curves become single hit, exponential functions of the particle flux density. This usually happens for mammalian cell survival curves also.

There are some cases, however, where even at the highest available LET exponentialization does not happen. Among these are survival of certain strains of diploid yeast, special strains of haploid yeast, tested while budding (37) and the radioresistant organism *Micrococcus radiodurans* (24). Diploid yeast was also shown to retain ability for post irradiation repair (38) following heavy ion exposure. If the major part of high LET radiation effects are due in these organisms to the cooperative effects and double chain scission occurs, then one could argue in these cases for the possible existence of "double chain repair mechanisms" in these instances. If, however, the pertinent genetic units are so small that the track core is not likely to be dominant in lethality, and single and coexistent lesions predominate, then ordinary single strand repair mechanisms may be sufficient. The repair (rejoining) of chromosome breaks is known to occur in various organisms: it seems to be important to continue investigations on the possibility that in diploid cells one complete double strand gene might act as template for enzymatic repair of a second set of genes which have double strand lesions. It is believed that in meiosis, while genetic recombination occurs, portions of double stranded genes detach from their own ploidy, freely mix and reconstitute. Somatic crossover in yeast is known to occur during postirradiation periods and can be regarded as a mode of repair for potentially lethal lesions. It seems that more research to find the exact molecular mechanisms involved might eventually lead to the recognition of molecular processes that are responsible for postirradiation repair of heavy ion lesions.

Carcinogenic (tumorigenic) action of radiation is known to occur with much increased efficiency at high LET and the increased carcinogenic hazard of fast neutrons and of alpha particle emitters is well recognized; the increased efficiency of high LET is sometimes attributed to the possibility that chromosome deletions and rearrangements must occur, and high LET radiation can cause such effects more efficiently than low LET. Recently H. H. Smith *et al.* (25) conducted some heavy ion experiments at the Berkeley Hilac with seeds of the plant arabidopsis. The heavy ions were found to have many times greater tumorigenic potency than gamma rays. It is of interest that, even with accelerated neon ions, tumorigenicity obeys the multiple hit dose effect relationship. Since one expects that a single neon ion is sufficient to kill or injure a single cell, one may wonder if the significance of multiple hit relationships in such instances might point to the possibility that sublethal injury to more than a single cell in a population of cells is necessary in some instances to disturb tissue regulatory mechanisms to such extent that tumor growth can be initiated. This idea is not entirely new; one of the great early medical physicists, G. Failla, proposed such a mechanism as early as 1936, and Austin Brues (40), who has carried out definitive experiments on the etiology of bone cancer, has deduced from his experiments that multicellular pathology usually precedes the appearance of radiation induced cancer. The availability of controlled beams of low and high LET particles should help us to make further quantitative contributions to this important problem.

REFERENCES

1. Lea, D. E., "Actions of Radiations on Living Cells," Cambridge University Press, Second Edition (1955).
2. Dessauer, F., "Quantenbiologie; Einfuhrung in einem neuem Wissenszweig", J. Springer Verlag (1964).
3. Pollard, E., "Primary Ionization as a Test of Molecular Organization", *Adv. Biol. Med. Phys.* **3,** pp 153-188 (1953).
4. Zirkle, R. E. and Tobias, C. A., "Effects of Ploidy and Linear Energy Transfer on Radiobiological Survival", *Arch. Biochem. Biophys.* **41,** pp 283 (1953).
5. Brustad, T., "Study of the Radiosensitivity of Dry Preparations of Lysosyme, Trypsin, and Desoxyribonuclease Exposed to Accelerated Nuclei of Hydrogen, Helium, Carbon, Oxygen, and Neon", *Rad. Res.* Supplement 2, pp 65-74 (1960).
6. Rossi, H. H., "Specification of Radiation Quality", *Rad. Res.* **10,** pp. 552 (1959).
7. Butts, J. J. and Katz, R., "Theory of RBE for Heavy Ion Bombardment of Dry Enzymes and Viruses", *Rad. Res.* **30,** pp. 855-871 (1967).
8. Baum, J. W., "Comparison of Distance and Energy Restricted Linear Energy Transfer for Heavy Particles with 0.25 to 100 MeV/amu", presented at Second International Congress on Medical Physics, Boston, 1969 (in press).
9. Turner, J. E., Hollister, M., "On the Relationship of Collision Theory to the Interpretation of Relative Biological Effectiveness", ORNL-4215, USAEC, Oak Ridge National Lab., pp 1-9 (January 1968).
10. Birge, A. C., and Sayeg, J. A., "The Effects of Accelerated Carbon Nuclei and other Radiations on the Survival of Haploid Yeast, Part I: Dosimetry of the Cyclotron Beams", UCRL-2293, Lawrence Radiation Laboratory (rev. 1958).
    Sayeg, J. A., Birge, A. C., Beam, C. A., and Tobias, C. A., "The Effects of Accelerated Carbon Nuclei and other Radiations on the Survival of Haploid Yeast, Part II: Biological Experiments", *Rad. Res.* **10,** pp 449-461 (1959).
11. Steward, P., "Stopping Power and Range for any Nucleus in the Specific Energy Interval 0.01 to 500 MeV/amu in any Nongaseous Material", Ph.D. thesis, University of California, Berkeley, UCRL-18127 (May 1968).
12. Litton, G. M., Lyman, J., and Tobias, C. A., "Penetration of High-Energy Heavy Ions, with the Inclusion of Coulomb, Nuclear, and other Stochastic Processes", Lawrence Radiation Laboratory UCRL-17392 (rev. May 1968).
13. Maccabee, H. D., Raju, M. R., and Tobias, C. A., "Fluctuations of Energy Loss by Heavy Charged Particles in Thin Absorbers", *Phys. Rev.,* **165,** No. 2, pp 469-474 (January 10, 1968).
14. Brustad, T., "Inactivation at Various Temperatures of the Esterase Activity of Dried Trypsin by Radiations of Different LET", *Rad. Res.,* Supplement 7, pp 74-86 (1967).
15. Tobias, C. A., and Todd, P. W., "Analysis of the Effects of High LET Radiations on Various Strains of Cells" In: *Biological Effects of Neutron and Proton Irradiations,* International Atomic Energy Agency, Vienna, Volume 2, pp 410-418 (1964).
16. Tobias, C. A., and Manney, T., "Some Molecular and Cellular Effects of Heavily Ionizing Radiations", Annals of the New York Academy of Sciences, **114,** Art. 1, pp 16-31 (March 31, 1964).
17. "Radiation Effects on Membranes", Proceedings of Symposium at Third International Biophysical Congress; Cambridge, Massachusetts, 1969; to be published in *Advances in Biological and Medical Physics,* Vol. 13, 1970.
18. Fluke, D. J., Brustad, T., and Birge, A. C., "Inactivation of Dry T_1 Bacteriophage by Helium Ions, Carbon Ions, and Oxygen Ions: Comparison of Effect for Tracks of Various Ion Density", *Rad. Res.* **13,** pp 788-808 (1960).
19. Dewey, D. L., "DNA Breaks in Colliphage T_7 after Irradiation with Argon Nuclei", *Int. J. Radiat. Biol.* **12,** No. 5, pp 497-503 (1967).
20. Powers, E. L., Lyman, J. T., and Tobias, C. A., "Some Effects of Accelerated Charged Particles on Bacterial Spores", *Int. J. Radiat. Biol.,* **14,** No. 4, pp 313-330 (1969).
21. Mortimer, R. K., Brustad, T., and Cormack, D. V., "Effectiveness of Ionizing Radiations for Induction of Mutations and Lethality in Diploid *Saccharomyces cerevisiae* in Relation to Ionization Density and Oxygen Tension", *Rad. Res.,* **26,** pp 465 (1965).

22. Todd, P. W., "Reversible and Irreversible Effects of Ionizing Radiations on the Reproductive Integrity of Mammalian Cells Cultured *in vitro*", Ph.D. thesis, Lawrence Radiation Laboratory, University of California, Berkeley, UCRL report 11614 (August 1964).
23. Barendsen, G. W., Walter, M. D. M., Fowler, J. F., and Bewley, D. K., "Effects of Different Ionizing Radiations on Human Cells in Tissue Culture: III, Experiments with Cyclotron Accelerated Alpha Particles and Deuterons", *Rad. Res.* **18,** pp 106 (1963).
24. Dewey, D. L., and Haynes, R. M., "Sensitivity of Micrococcus Radiodurons to Densely Ionizing Radiations", UCRL 17481, Lawrence Radiation Laboratory, Berkeley, pp 138-144 (Spring 1967).
25. Hirono, Y., Smith, H. H., and Lyman, J. T., "Tumor Induction by Heavy Ionizing Particles and X-Rays in Arabidopsis", *Rad. Botany* **8,** pp 449-456 (1968).
26. Rosenzweig, W., and Rossi, H. H., "Statistical Fluctuations in the Energy Loss of 5.8 MeV Alpha Particles Passing Through an Absorber of Variable Thickness", NYU-10716, Columbia Radiological Research Laboratory Report, (November 1963), unpublished.
27. Henriksen, T., "Free Radicals Induced in Enzymes by Electrons and Heavy Ions", *Rad. Res.* Supplement 7, pp 87-101 (1967).
28. Horan, P., "Radiation-Induced Free Radical Destruction", Ph.D. Thesis, Pennsylvania State University, and Lawrence Radiation Laboratory (June 1970).
29. Cory, P. M. and Cole, A., "Radiation-Induced Double-Strand Scission of the DNA of Mammalian Metaphase Chromosomes", *Rad. Res.* **36,** pp 528-543 (1968).
30. Elkind, M. M., "Damage Registration and Repair Following Neutron Irradiation" (in press).
31. Manney, T. R., Brustad, T., and Tobias, C. A., "Effects of Glycerol and of Anoxia on the Radiosensitivity of Haploid Yeasts to Densely Ionizing Particles", *Rad. Res.* **18,** pp 374-388 (March 1963).
32. Curtis, S. B., "Human Kidney Cell Oxygen Enchancement Ratios for Fast Neutron Beams and a Prediction for Negative Pion Beams", UCRL-19456, Lawrence Radiation Laboratory (January 1970).
33. Norman, A., "Thermal Spike Effects in Heavy-Ion Tracks", *Rad. Res.* Supplement 7, pp 33-37 (1967).
34. Oda, N. "Relationship Between Primary Energy Transfer and Target Theory", in: *Biophysical Aspects of Radiation Quality,* IAEA, Vienna, pp 165-183 (July 1966).
35. Fleischer, R. L., Price, P. B., and Walker, R. M., "Tracks of Cosmic Rays in Plastics", *Science* **13,** Vol 155, pp 187 (January 1967).
36. Benton, E. V., "A Study of Charged Particle Tracks in Cellulose Nitrate", Ph.D. thesis, Stanford University (March 1968).
37. Elkind, M., and Beam, C. A., "Variation of the Biological Effectiveness of X-Rays and ε-particles on Haploid *Saccharomyces cerevisiae*", *Rad. Res.* **3,** No. 1, pp 88-104 (September 1955).
38. Lyman, J. T., and Haynes, R. H., "Recovery of Yeast after Exposure to Densely Ionizing Radiation", *Rad. Res.* Supplement 7, pp 222-230 (1967).
39. Oda, N., "Calculation of Primary Energy Transfer and Effective Inactivation Cross Section of Targets for Heavy Charged Particles Based on Microdose Concept", JAERI 1179, Japan Atomic Energy Research Institute, (1969); and "Calculations of Survival Rates and Effective Inactivation Cross Section of Targets for Heavy Charged Particles Based on Microdose Concept", JAERI 1180 (1969).
40. Brues, A., "Carcinogenic Effects of Radiation", in: *Advances in Biological and Medical Physics,* Volume 2, edited by John H. Lawrence and Joseph G. Hamilton, Academic Press, Inc., New York, 1951.

GALVANIC CELL TRANSDUCERS FOR THE IN VIVO MEASUREMENT OF OXYGEN TENSION

John S. Clifton
Dawood Parker
Alfred C. C. Tseung

Physiological Considerations

Inspired air contains 20.9% oxygen representing a partial pressure of 159 mm Hg. As this air enters the alveoli it is diluted by carbon dioxide and water vapour so that the final partial pressure of oxygen falls to about 104 mm Hg. The oxygen passes by diffusion through the walls of the alveoli and lung capillaries into the blood stream. The partial pressure of oxygen in blood leaving the lungs is usually between 90-100 mm Hg. Oxygen given up from the haemoglobin in exchange for carbon dioxide diffuses through the capillary walls into the cells. As a consequence the blood returning to the heart via the veins has a partial pressure of about 40 mm Hg.

Various diseases affect the amount of oxygen transferred from the alveoli to the blood, the amount carried in the blood, and the amount extracted and utilised by the tissues. Knowledge of blood oxygen tension can therefore yield very valuable information for the investigation and diagnosis of many disease processes. For example, by measuring the oxygen content of both the arterial and venous blood samples from a given area and calculating the A-V difference, an indication of the oxygen extraction by the tissues in that particular area can be obtained. However, this will indicate only the amount of oxygen in the blood at capillary level. More significant is a knowledge of the oxygen actually available to the cells. This can only be obtained by a direct measurement in the tissue concerned.

Current Methods of Measuring pO_2

The physical properties characteristic of oxygen which have been utilised in its detection are its high magnetic susceptibility and its selective light absorption when combined with haemoglobin in the blood.

The paramagnetic susceptibility of oxygen is used in industrial applications to monitor oxygen in gas streams. A relatively large volume of oxygen is required for such analysis and the response time of such magnetic oxygen analysers is slow.

The large difference between the optical absorption coefficient of haemoglobin and oxyhaemoglobin in the red region of the spectrum is used in transmission and reflection oximeters for the determination of blood oxygen saturation. This method does not give a direct determination of blood oxygen tension.

Physico-chemical methods of detecting oxygen which are of importance are gas chromatography, mass spectrometry and polarography. By either gas chromatography or mass spectrometry an analysis is possible of virtually all the components present in clinical mixtures of anaesthetic and respiratory gases.

With the exception of polarography, all these methods of determining oxygen tension are essentially in vitro techniques requiring relatively large volume samples and are complex in nature. The ear oximeter and cuvette type oximeter have been used as continuously recording devices but the measurements so obtained are of blood oxygen saturation and not oxygen tension.

There has been increasing interest in recent years in *in vivo* continuous recording transducers for blood oxygen tension. In this application where simplicity, sterility and small size are essential,. the polarographic method of Clark (1956) for determining oxygen tension has offered to date the most practical and commonly used technique.

The polarographic oxygen electrode depends for its operation on the irreversible decomposition of oxygen in the presence of an electrolyte at the surface of a noble metal cathode held at an applied negative potential. This negative potential is so adjusted that total reduction takes place of all the oxygen reaching the cathode. A gas permeable membrane is used to isolate the electrode from its environment. The current flow from between cathode and anode in this case is a direct measure of the oxygen tension around the cathode.

However, irreversible chemical reactions can take place in these electrodes thus greatly reducing their operating life. Long term instability due to leakage currents associated with insulation breakdown between cathode and anode can also be a problem with these electrodes.

Further the measurement of small currents from high impedance sources necessitates complicated electronics. In an attempt to overcome these difficulties we have directed our attention to the electrochemical determination of oxygen tension by the galvanometric technique. The resulting galvanic oxygen cell is a simple, robust, low-impedance device, free from the problem of insulation breakdown associated with high-impedance polarographic electrodes. Since the cell generates its own power, an external source of applied voltage is not required.

Galvanic Oxygen Cells

Development

Although the use of galvanic cells (often called metal/air cells) to derive electrical energy from atmospheric oxygen has been known for some time (Grove 1839) the use of the current output from such a cell as an indicator of oxygen tension in a surrounding fluid has been little exploited.

A galvanic oxygen cell for measuring oxygen partial pressure in gas streams was first reported by Paris (1933) who used a zinc-carbon cell. Devices based on the galvanic principle have been used to monitor oxygen tension in water, rates of oxygen uptake in biological systems and to determine oxygen concentration in industrial processes (Hoare 1968).

One of the earlier devices developed at U.C.H. was a "flow-through" silver-lead galvanic cell isolated from the test fluid by a polytetrafluoroethylene (P.T.F.E.) membrane, which measured continuously the oxygen tension of blood passing from a heart-lung machine to patients undergoing heart surgery. (Figure 1)

Progressive development of the basic cell system has resulted in a direct and continuously measuring pO_2 device mounted in the tip of a cardiac catheter which permits the in vivo recording of oxygen tension. Some areas of research in which this device is currently being evaluated are

> Cardiac surgery where continuous monitoring of arterial and venous blood provides a direct indication of patient condition, and by A-V difference, an estimation of cardiac output.
>
> Resuscitation under hyperbaric oxygen of asphyxiated new born infants.
>
> Control of patients with rhesus antibodies where the oxygen tension of amniotic fluid is thought to reflect foetal oxygenation and well being.

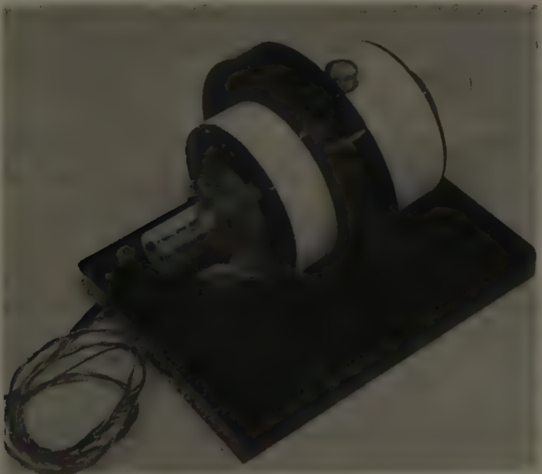


Fig. 1. "Flow-Through" Galvanic Cell for pO_2 measurements in Open Heart Surgery.

Theory

The distinctive features of galvanic cells used for the determination of oxygen tension are:—

an 'in-built' electrochemical potential difference across the cell;
a low cell impedance;
a cell life which depends on the current drawn and the amount of anode material available for the cell reaction.

When a noble metal cathode is placed in an electrolyte containing oxygen a voltage develops in the half cell so formed. The origin of this voltage can be explained in the following manner.

Any noble metal will act as a catalyst in the reduction of oxygen. Therefore, when one of these metals is placed in an electrolyte in the presence of oxygen, the reduction of oxygen takes place.

$$\tfrac{1}{2}O_2 + H_2O + 2e^- \rightarrow OH^-$$

The result is that the cathode acquires a positive potential.

The electrons to maintain the reaction must however be supplied by a reference electrode.

Using a standard hydrogen electrode as reference, the potential developed across this cell is 1.23 V (Latimer, 1952). When this reference electrode is replaced by a metal anode e.g. lead, the theoretical open circuit voltage drops to 0.95 volts (Latimer, 1952). This open circuit voltage does not vary linearly with oxygen tension so that the cell has to be polarized, by the use of a load resistor, to operate in a region where its current output is directly related to oxygen tension.

A schematic diagram of a typical galvanic oxygen cell is shown in Figure 2.

The silver cathode and lead anode are immersed in a potassium hydroxide solution and connected via an ammeter of internal resistance R_L. As a result a current flows from the silver to the lead, the magnitude of which is dependent on the rate at which dissolved oxygen in the electrolyte diffuses towards the silver cathode. The origin of this current and its dependence on the rate of supply of oxygen to the silver cathode are explained by the following reactions which occur at the cathode and anode surfaces.

Since lead is more electronegative than silver the tendency is for lead to go into solution with the formation of lead ions on the lead surface accompanied by a flow of electrons from the lead to the silver.

$$Pb \rightarrow Pb^{2+} + 2e^-$$

When oxygen is present at the surface of the silver electrode, it reacts with the electrons to produce hydroxyl ions.

Cathodic reaction $\tfrac{1}{2}O_2 + H_2O + 2e^- \rightarrow 2OH^-$

At the anode the lead ions combine with hydroxyl ions as follows:—

Anodic reaction $Pb + 2OH^- \rightarrow Pb(OH)_2 + 2e^-$
Overall cell reaction $Pb + H_2O + \tfrac{1}{2}O_2 \rightarrow Pb(OH)_2$

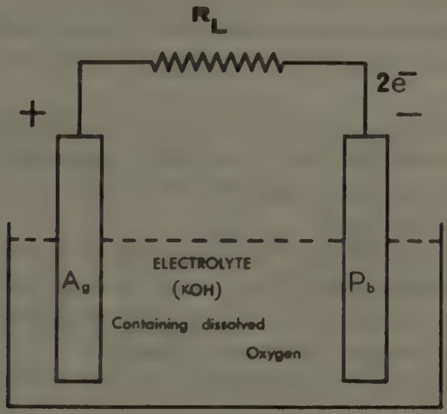

R_L
2e⁻
+
−
Ag
Pb
ELECTROLYTE
(KOH)
Containing dissolved
Oxygen

Fig. 2. Schematic Diagram of a Theoretical Silver-Lead Oxygen Cell.

The current flowing in the external circuit is a measure of the rate of this reaction and thus of the rate of supply of oxygen to the silver electrode.

In the construction of devices for measuring oxygen tension, oxygen is made to diffuse through a gas permeable membrane which is in close contact with the cathode. Such a cell assembly, isolated from the test fluid by a diffusion membrane, is shown in Figure 3.

For a particular membrane of constant permeability the rate of supply of oxygen to the cathode depends on the oxygen tension gradient across the membrane. Since the oxygen concentration inside the cell is maintained at a constant

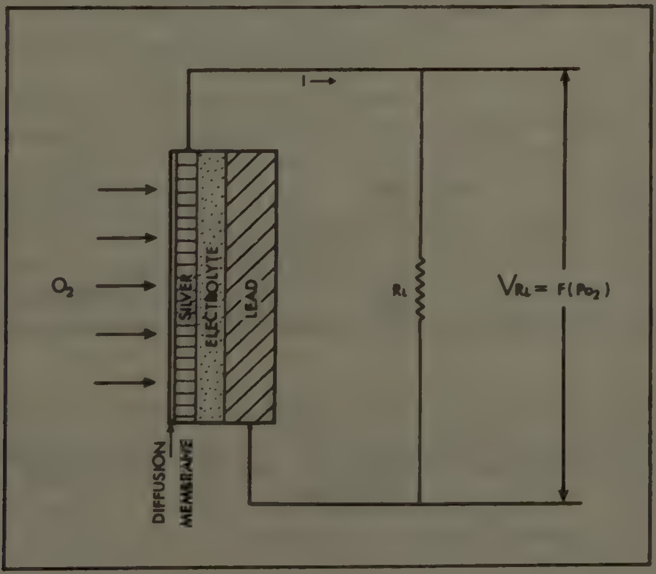

I →
O₂
SILVER
ELECTROLYTE
LEAD
DIFFUSION MEMBRANE
R_L
V_RL = F(Po₂)

Fig. 3. Schematic Diagram of an Oxygen Cell incorporating a Diffusion Membrane.

low level by the cell reaction, the oxygen tension gradient across the membrane is a function of the external oxygen tension. The current flowing in the external circuit is therefore directly related to the oxygen tension in the fluid to which the membrane is exposed.

The effect of applying a load to the cell is to change the potential of both cathode and anode i.e. the electrodes polarize in opposite directions. The cathode becomes less cathodic and the anode less anodic thus decreasing the available cell voltage. This is shown in Figure 4.

Graphs show the variation of the potentials of silver and lead with current density, in an air saturated electrolyte. It can be seen that the silver polarizes in three distinct steps and reaches a limiting current density which is dependent only on the concentration of oxygen around the cathode.

Figure 5 shows the variation of limiting current with oxygen tension for a silver-lead cell.

It can be seen that the limiting current is linearly related to the pO_2 except at very high values of pO_2 where a slight deviation occurs. From this it is clear that to ensure the successful operation of a galvanic cell for the determination of oxygen tension the load resistance must be chosen such that the cell operates in its limiting current region. For most of our devices a load of 100 ohms has been suitable.

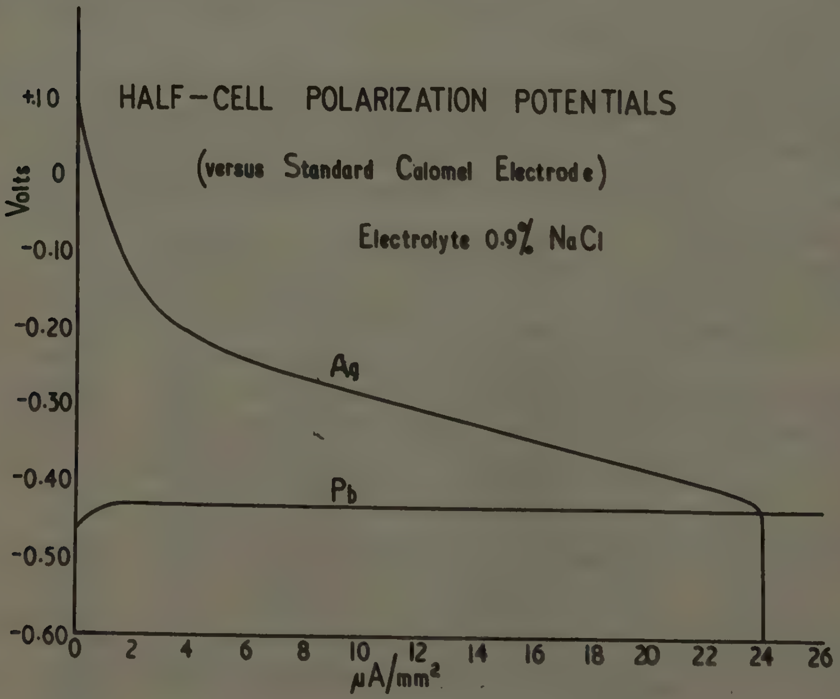


Fig. 4. Graphs showing relationship of the Half Cell Potentials of Silver and Lead with respect to current density in an air saturated electrolyte.

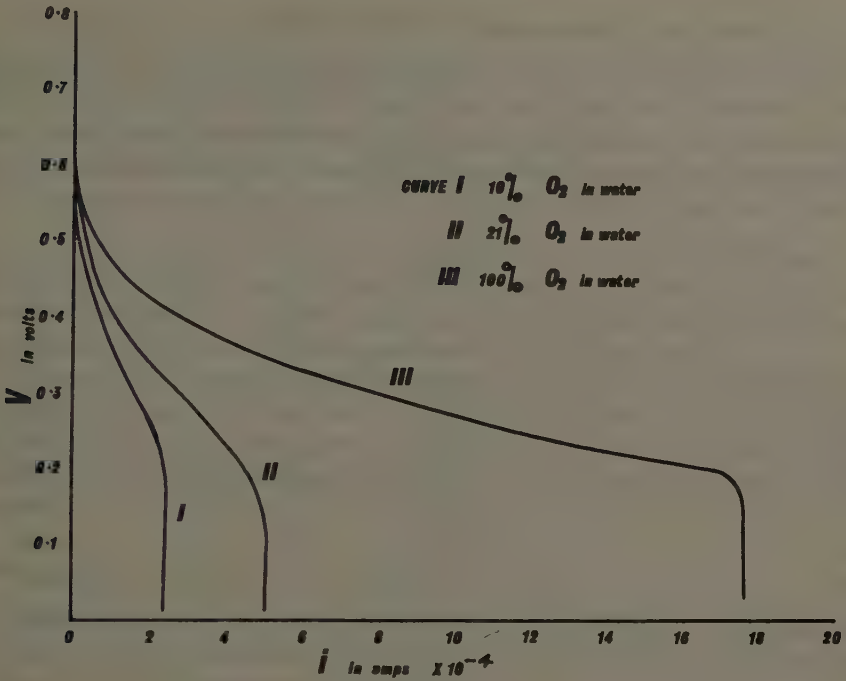


Fig. 5. Current-Voltage characteristic for a typical Silver-Lead Oxygen Cell. Limiting cell current is proportional to oxygen tension.

Choice of Materials

Anode Materials: The main function of the anode is to provide a source of electrons to the cathode. To ensure long term stability of the current output from the cell, the anode should have certain properties.

> The requirements for an efficient anode are
> an electro-chemical potential sufficiently more negative than that of the cathode, to maintain self polarization of the cell:
> a low rate of self-discharge:
> the ability to maintain a constant output current for long periods i.e. the oxide or hydroxide films formed on the anode surface should not cause heavy polarization of the anode:
> for in vivo use the products formed in the cell reaction should not be toxic.

The potential generated by oxygen reduction at the cell cathode is high i.e. +1.23 V (measured against a standard hydrogen electrode). The first requirement listed above is, therefore, satisfied by a number of non-noble base metals. However, few metals satisfy the remaining requirements for an efficient anode material.

The most common form of self-discharge of metal anodes is due to hydrogen evolution.

$$2M + 2H_2O \rightarrow 2MOH + H_2$$

The work of Pourbaix et al. (1966) provides a useful guide for the selection of the best combination of electrolyte and anode material. For a particular metal anode they give the pH range of the electrolyte which would ensure a negligible rate of hydrogen evolution and therefore a low rate of self-discharge of the anode. Their results are summarized in Table I.

Table I

	pH Electrolyte Range for Low Self-Discharge	*Current/g*
Al	4–9	2.98
Be	4–10	5.94
Sn	0–10	0.90
Fe	2–14	1.43
Pb	8–11	0.81
Ni	8–10	0.91
Co	8–10	0.90
Cd	9–11	0.47

All the metals in Table I, with the exception of Beryllium which is toxic, can be used as an anode material. Aluminium appears to be the best choice because of low self-discharge over a wide pH range and its high current/g value. In the determination of oxygen tension, where a high cell current is not important, lead in an alkali solution is also a good choice.

Cathode Materials: The basic requirements of a galvanic cell cathode are:

it should be able to reduce oxygen efficiently;
it should be corrosion resistant.

Work on fuel cell oxygen electrocatalysts has shown that Platinum, Palladium and Silver are all efficient for reducing oxygen and are corrosion resistant. Silver was chosen because of its low cost.

Electrolytes: The use of galvanic cells to determine oxygen tension in fluid originated from work on metal-air batteries where KOH is used extensively as the electrolyte because of its high conductivity. In galvanic oxygen cells KOH is not always a good choice since even 1N KOH will have a pH greater than 12. This means that most of the anodes listed above will not be operating in their optimum pH range and the evolution of hydrogen becomes possible. A further disadvantage is the possible deposition on the anode of a carbonate formed by the reaction of KOH and carbon dioxide in the air. The best choice of electrolyte is a near neutral solution such as 0.9% NaCl or sodium sulphite solution. With these electrolytes the likelihood of hydrogen evolution is negligible.

Membranes: Any material which is gas permeable, but impervious to ions, can be used as a membrane. However, for devices to be used in vivo the membrane must satisfy the following requirements:—

ensure that oxygen can diffuse to the cathode of the cell;
provide protection for the cell from direct contamination by body fluids;
not promote coagulation when used in blood.

The two materials which best meet these requirements are Teflon and silicone rubber. There are technical difficulties with the use of Teflon since the surface must be etched before an adhesive can be applied. Silicone rubber membranes can be applied by dip coating techniques and cured at low temperatures.

The response time of the cell is in general proportional to the thickness of the membrane. However, for very thin membranes cell response is limited by thickness of electrolyte between membrane and cathode. A Teflon membrane of 6 micron or a silicone rubber membrane of 25 micron will give a response time of about 1 second.

Devices for In Vivo Use

Using the principles previously described, galvanic oxygen cells suitable for in vivo use have been built into the tip of catheters of different sizes. Three distinct types of catheter cells have been developed.

A catheter-tip oxygen cell in which the nylon catheter wall acts as the diffusion membrane.
A catheter-tip cell with a removable silicone rubber cap membrane which is in intimate contact with the anode and cathode in the catheter tip.
A catheter-tip cell with a dip-coated membrane tip of silicone rubber.

Figure 6 shows the cell assembly used in the first type of catheter. The catheter wall acts as the oxygen diffusion barrier. The anode is a thin lead strip wrapped tightly round an insulated wire which also serves as the cathode contact wire. The cathode is a silver wire coil. A second insulated wire is soldered to the far end of the lead strip. These two very fine insulated wires go through the catheter and out to the external circuit.

Electrolyte (both 0.9% NaCl and 1N KOH have been used) is sucked up the catheter tube, care being taken to avoid the formation of air bubbles in the cell. The rear of the cell, including the electrical contacts, is then sealed off with dental wax, the wax being melted in position by application of external heat. Care is taken to insulate the soldered contacts from the electrolyte to prevent the occurrence of auxiliary electro-chemical cells within the catheter. The tip of the catheter is finally sealed with epoxy resin and rounded off.

Typical catheters made from Teflon and nylon are illustrated in Figure 7.

The larger Teflon catheter is of 1.5 mm external diameter with a wall thickness at the catheter tip of 0.01 mm. The linear variation of cell output with pO_2 at 37°C is 2.1×10^{-7} amp per % oxygen. The response time of the cell to 98%

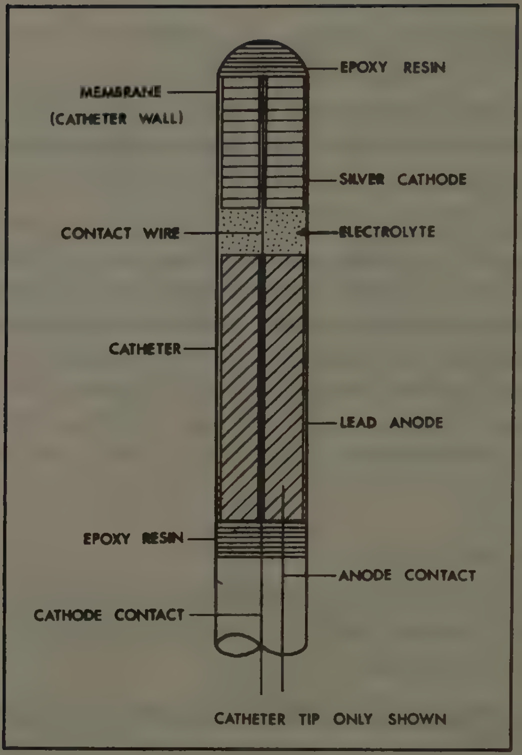

EPOXY RESIN
MEMBRANE (CATHETER WALL)
SILVER CATHODE
CONTACT WIRE
ELECTROLYTE
CATHETER
LEAD ANODE
EPOXY RESIN
ANODE CONTACT
CATHODE CONTACT
CATHETER TIP ONLY SHOWN

Fig. 6. Section through an "In Vivo" Catheter Oxygen Cell.

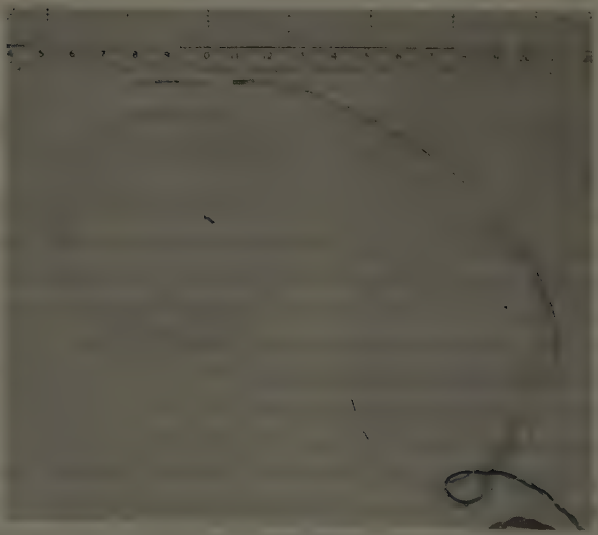


Fig. 7. Examples of "In Vivo" Catheter Oxygen Cells. Scale in millimeters.

change of oxygen concentration at 37°C is 3 minutes. The temperature sensitivity of the cell is 3% per °C.

The galvanic cell has also been incorporated into a nylon flow-catheter of external diameter 0.63 mm, wall thickness 0.06 mm. After the incorporation of the cell into its tip, the catheter retains its flow properties and flexibility. At 37°C the cell output is 1.2×10^{-8} amp per % oxygen. (The cell load resistance was 100 ohms). The response time at 37°C is 2 minutes. Cell output was unchanged after continuous operation in air-saturated saline at 37°C for 18 hours. A conservative estimation of the life of these cells is 50 hours of continuous use in air.

The two disadvantages with this type of cell construction are its slow response, and limited lifetime. Once the cell has lost its oxygen sensitivity due to the conversion of the lead anode to lead hydroxide there is no means of restoring this sensitivity.

Figure 8 shows the second type of catheter oxygen cell. A cathode of 150 micron diameter silver wire is insulated from the lead anode by epoxy resin. The outside diameter of the nylon catheter is 1.5 mm.

The method of assembly is to wet the catheter tip containing the anode and cathode with electrolyte. A 0.9% solution of NaCl is used. The tight fitting silicone rubber cap is then slipped over the catheter tip and sealed onto the catheter with silicone rubber adhesive. The thickness of the rubber cap at the catheter tip is about 25 microns.

The current output from such a cell depends on the membrane thickness and the temperature. A typical value for the current output is 1×10^{-9} amp per % oxygen at 37°C. The response time of the cell in this case is 3 seconds.

A further advantage of this cell arrangement is the fact that at the end of the cell life the membrane can be removed and the cell reactivated by removing the hydroxide film from the catheter tip by scraping or by dipping in a 10% solution of acetic acid.

Both types of catheter cell can be sterilized in 10% ethylene oxide gas.

The third type of catheter cell is a recent development. The geometry of the cell is as shown in Figure 8. The catheter tip is moistened with electrolyte and the

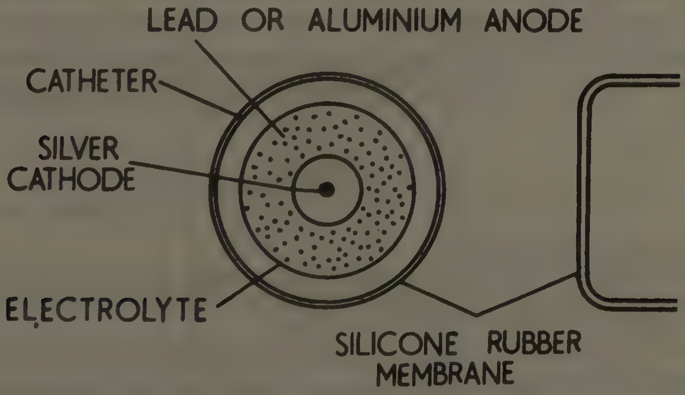


Fig. 8. End Section of a Catheter with replaceable Silicone Rubber Membrane.

membrane applied by dip coating the tip with silicone rubber. This dip coating technique markedly simplifies construction. Further, because of the minimal electrolyte film between cathode and dip-coated membrane the response time is improved. Response times of less than a second at body temperature have been achieved by this technique.

It is also found that water lost by evaporation through the membrane during storage can be replaced by holding the catheter tip in water vapour. The possibility of using the electrolyte in dry crystal form before dip-coating is therefore under investigation. In this case the cell could be activated by introducing vapour through the membrane after the dip-coating process. A particular advantage of this type of cell would be the possibility of activating the cell during steam sterilisation.

Calibration

Linearity of Response

The linearity of the response of these galvanic cells for different oxygen tensions was checked in saline solutions containing 2, 5, 10, 21 and 100% oxygen. The zero of the calibration was determined by using pure nitrogen.

However, a zero oxygen point is not usually required for calibrating the cell output. The cell can be satisfactorily calibrated against only one known oxygen tension, the simplest of which is air-saturated saline. For zero oxygen tension zero current may be assumed. For thick Teflon and nylon membranes as used in the 'flow-through' and first type of catheter cell described, the cell output current was linear with oxygen tension in the range 0-710 mm Hg. For thin Teflon and silicone membranes the response was linear up to an oxygen tension of 250 mm Hg. Above this oxygen tension the cathode is unable to reduce all the oxygen diffusing into the cell and the output current reaches a limiting value.

Flow-rate Dependence

A limiting factor in the use of these galvanic cells is their consumption of oxygen. The layer of fluid immediately surrounding the membrane is rapidly depleted of oxygen. In the absence of fluid flow the current output from the cell would decrease as the oxygen around the cathode is consumed. The reduction of oxygen tension, $-\Delta pO_2$, around the cathode is given by (Gleichmann and Lübbers, 1960):

$$-\Delta pO_2 = \frac{2.66 \times 10^3 i}{\alpha V} \qquad \text{mm Hg/min}$$

where i = cell current in amps
α = the solubility of oxygen in the fluid in ml of O_2/ml at 760 mm Hg.
V = volume of solution in ml.

From this equation it is clear that the cells reported here are not suitable for the determination of oxygen tension in static fluid. For the satisfactory operation of these cells in vivo a flow-rate of at least 10 cm/sec is required.

Temperature Dependence

At constant oxygen tension the cell current i_T at $T°K$ increases with temperature (Mancy et al, 1962)

$$i_T = K \exp(-J/T)$$

where K and J are constants for a particular cell geometry. For Teflon membranes the temperature coefficient is about 3% per °C. For silicone rubber an average value of 2% per °C is obtained.

The galvanic oxygen cell used on the heart lung machine was operated in conjunction with a temperature compensation circuit consisting of a wheatstone bridge with thermistors in opposite arms of the circuit (Briggs and Vincy, 1964). In this way temperature compensation was obtained to within 2% in the temperature range 10–40°C, permitting satisfactory operation under conditions of hypothermia.

For in vivo use no temperature compensation is required if the calibration of the cell is made at blood temperature.

Work in Progress

As has been pointed out before, for very thin, highly permeable membranes the response time of the cell is determined not by the thickness of the membrane but by the thickness of the electrolyte layer between the membrane and the cathode. This limitation poses a problem in such applications as breath by breath monitoring of oxygen partial pressure where a response time of less than a second is required.

Work in progress concerns the development of a silver cathode evaporated onto the diffusion membrane. The membrane is a film of Teflon 6 microns thick and held in position by an O-ring. The anode is a lead cylinder. The cell is shown in Figure 9.

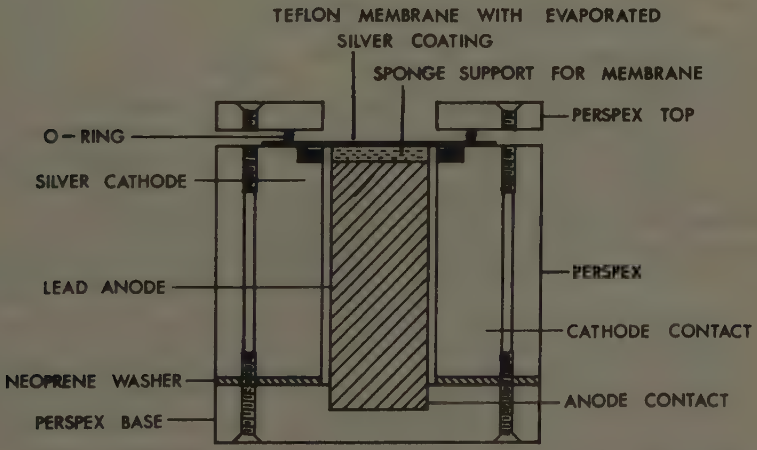

TEFLON MEMBRANE WITH EVAPORATED SILVER COATING
SPONGE SUPPORT FOR MEMBRANE
PERSPEX TOP
O-RING
SILVER CATHODE
LEAD ANODE
PERSPEX
CATHODE CONTACT
NEOPRENE WASHER
ANODE CONTACT
PERSPEX BASE

Fig. 9. Section of a "Fast-Response" Oxygen Cell with evaporated Silver Cathode.

With this technique the electrolyte/metal interface can, by evaporation of a suitable film thickness, be situated at less than 0.01 micron from the membrane. Response times of 0.30 seconds (98% change) have been achieved at room temperature using this technique.

The cell is so constructed that the membrane-cathode film is disposable and the anode can be cleaned and removed when necessary.

To enable the galvanic cell technique to be used in static fluid, work is currently in progress on the development of a cell with a 1 micron diameter silver cathode. With a cathode of this size the oxygen consumption by the cell will be negligible (Hudson, 1967) and consequently the cell current will not be dependent on flow-rate.

Conclusion

The galvanic oxygen cell, as developed, provides a simple, robust, and reliable device for the in vivo measurement of oxygen tension. Its low impedance, ease and reproducibility of calibration offer advantages over the more commonly used polarographic technique.

Present developments indicate that cells for use under static conditions, cells with very fast response times, and cells that are pre-sterilized and disposable can be achieved.

Acknowledgement

The authors gratefully acknowledge the financial support given to this work by the British Empire Cancer Campaign, and latterly the Medical Research Council.

REFERENCES

Briggs, R., and Vincy, M., 1964, *J. Sci. Instrum.*, **41,** 78.
Clark, L. C., 1956 *Trans. Am. Soc. Art. Int. Org.* **25,** 41.
Gleichmann, U., and Lübbers, D. W., 1960, *Pflügers Arch.* **271,** 431.
Grove, W. R., 1839, *Phil. Mag.*, **3**, 14, 139.
Hoare, J. P., 1968, The Electrochemistry of Oxygen, Interscience Publishers.
Hudson, J. A., 1967, *Med. and Biol. Engng.*, **5,** 207.
Latimer, W. M., 1952, The oxidation states of the elements and their potentials in aqueous solutions 2nd Edition, Prentice-Hall.
Mancy, K. H., Okun, D. A., and Reilley, C., 1962, *J. Electroanal. Chem.*, **4,** 65.
Paris, A., 1933, *Ind. Chim.*, **20,** 807.
Parker, D., and Clifton, J. S., 1967, *Phys. Med. Biol.*, **12,** 2, 193.
Pourbaix, M., 1966, Atlas of Electrochemical Equilibria in Aqueous Solutions, Pergamon Press.

MEASUREMENT OF HUMAN PHYSIOLOGICAL RESPONSES

H. S. Wolff

EVEN GOOD TECHNOLOGY IS NOT ENOUGH

Terms such as physiological monitoring, intensive care monitoring, and patient monitoring have been part of the medical instrumentation language since the late 1950's. During the following six years, many firms in part stimulated by astronaut instrumentation contracts, started to manufacture patient monitoring equipment. With a few exceptions these endeavours proved not only financially unrewarding to the firms, but also of little use to the hospital. It is only comparatively recently, particularly in the field of coronary care equipment that commercially viable, and useful equipment has come onto the market. On a purely intellectual analysis, the use of monitoring could be justified on two general grounds:

1. A seriously ill patient can be considered as one in which one or more of the homeostatic systems which keep his internal environment within limits compatible with life, is either no longer operative, or has changed its "set" point in an unfavourable direction.

   This means that control has to be exercised by an external agency, through the administration of drugs, replacement of body fluids, control of temperature, electrical stimulation etc.

   For the control to be effective there has to be feed-back of the value to be controlled to the administrator of the therapy. Often this feed-back is obvious to the patient himself, who just feels better, or to the nursing and medical staff, who make use of easily elicited clinical signs. If the physiological information is troublesome or tedious to acquire, or if the rate of change can be so fast as to require very frequent measurement, then it becomes reasonable to acquire some of it by automatic means. This process, with the possible addition of alarms, is the basis of patient monitoring.
2. Under certain conditions physiological signals can be acquired in advance of major systemic changes. These may be electrical or chemical in nature, and afford the opportunity to anticipate a crisis. The best example is an increase in ventricular extra-systoles, as indicated by the electro-cardiogram, which in themselves will only have a small effect on circulatory parameters, but can be taken to be an indication of an increased probability of subsequent ventricular fibrillation.

Definition

First, a problem in semantics. Equipment that is capable of making a physiological measurement on a patient and displaying the result of such a procedure is not necessarily a *monitoring* system, in the sense in which the term will be used in this paper. To qualify for the term monitoring, the equipment would satisfy the following criteria:

1. It must have the basic capability of attracting the attention of someone who is *not actually watching* the output end of the system when a given measurement falls outside preset limits. Whether this is done by actual alarms or other means: it implies that the equipment actively communicates with the staff rather than that it waits to be interrogated.
2. It should produce permanent records of such a nature that not only the actual values of past measurements, but also their trend can readily be envisaged. These records may be of two kinds:
    a. A time series extending over hours or days of the values which change comparatively slowly, for example the heart rate, temperature and blood pressure. This record may be complementary to, or used instead of charts kept by nurses (Figure 1).
    b. Records of phasic events, such as bio-electric waveforms, either on paper (in which case only intermittent samples can conveniently be kept), or on magnetic tape, which does allow continuous recording and subsequent re-visualization.

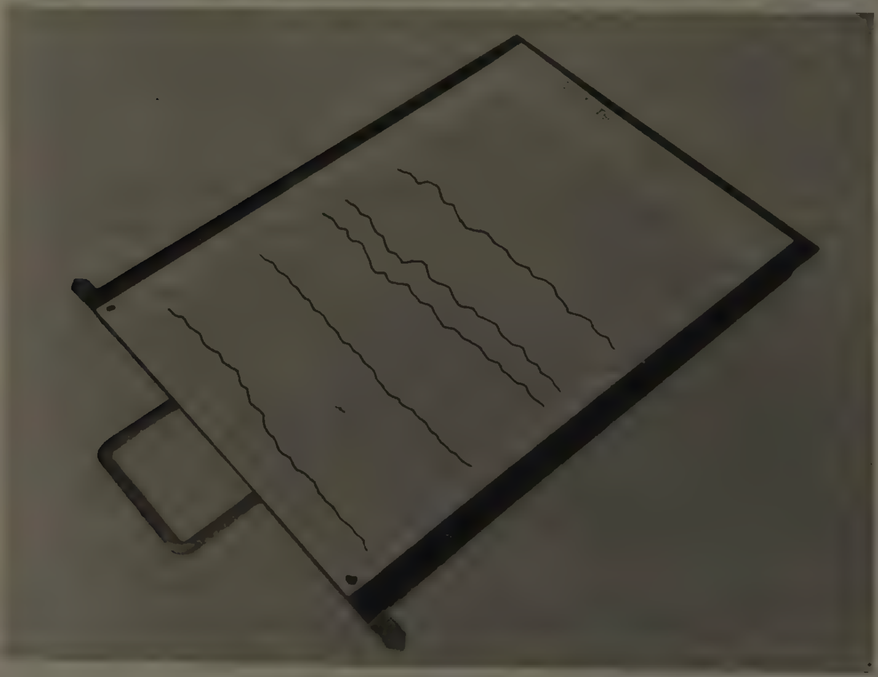


Fig. 1. Chart from automatic five-channel printer. The chart is visible in the printer, and can be withdrawn from it and reinserted at any time.

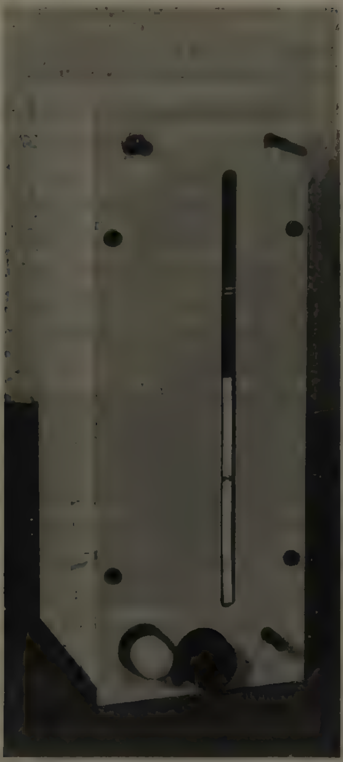


Fig. 2. Single channel from indicator bank. The length of the illuminated column is proportional to the magnitude of the measurement. The bars across the column indicate preset upper and lower alarm limits.

3. A method of display of the measured values which:
   a. indicates clearly the type of measurement.
   b. Shows the instantaneous value in the form in which it would actually be quoted or reported: for example, if the arterial blood pressure is going to be quoted as 120/90, then the figures 120 and 90 must be easily relatable to the indication (Figure 2).
   c. gives the relationship of the indicated value to the safe upper and lower limits, which may have been set for a particular patient. This is to avoid the necessity for the staff to engage continually in minor mental computations in which the indicated value is compared with a remembered normal or limiting value, before a decision for action is taken on the basis of the magnitude of the difference.
4. To satisfy 2(a), 3(a), 3(b) and 3(c), at least a minimum of data processing must be supplied. As a simple example, the display of a number of EKG complexes on an oscilloscope does not itself readily yield a figure of heart rate, though of course, the spacing of the complexes for a given time-base speed, gives some indication. To drive an indicator and/or alarm system, processing by a rate circuit is required.

Equipment which does not at least satisfy a substantial portion of these criteria, I should prefer to call "a physiological measuring system" and not "a physiological monitoring system". This distinction does not in itself imply any superiority of one over the other. What is important is to realise that the way in which the two systems are employed; the attitude of the staff to the system, and the staff requirements may well be radically different.

Broadly speaking, a physiological measuring system for a given parameter will be technically identical, as far as transducers and amplifiers are concerned, to a full monitoring system. From then onwards it will be less complex and cheaper, but it will have to rely on the constant availability and presence of vigilant and motivated staff for the interpretation of the indications, for the detection of dangerous trends and for the keeping of records.

Conversely, a monitoring system will be more complex and more expensive, but it will be much less dependent on a high staff/patient ratio, and if trend and alarm indications are used intelligently, will make smaller demands on vigilance and motivation. Which system is appropriate depends on the local situation. In a teaching hospital, with a pool of "doctors" in training and no desperate shortage of nurses, a physiological measuring system may well be appropriate.

I am, however, firmly convinced that thinking on a national scale, technological innovation should be used as a leveller, to make standards of care and treatment possible in the less fortunate hospital which previously could only be achieved for a minority of patients in a "good" establishment. To achieve this levelling action, equipment must make the minimum of extra demands on the staff, and this implies substituting technical sophistication not only for the *mere presence of human beings*, but also for intelligence and motivation. It also implies a standard of reliability, which can be achieved with the best possible engineering, but which has not, as yet, been applied, largely for reasons of cost and design inexperience.

Why does one monitor in intensive care?

A basic assumption behind all intensive care monitoring is that the rate of change of the physiological state may be sufficiently high to warrant continuous, or near-continuous data acquisition. If, therefore, measurements were to be made by conventional methods, where these are available, the required frequency might overload the staff. One therefore looks for automatic data acquisition by means of transducers in continuous contact with the patient.

Because data is continuously available, it is possible to detect unfavourable trends sooner, which in turn allows counter-measures to be taken before the situation becomes really dangerous. The effects of "quick action" type therapy can be seen immediately by the person applying the therapy without at the same time loading him with a measuring procedure. If vigilance is relaxed, the supervisory alarm systems will draw attention to a potentially dangerous situation. It would, however, be wrong to regard a monitoring system primarily as a "fire alarm" to indicate that an emergency had occurred. Instead, it should be so organized by the choice of suitable clinical parameters, that the emergency will be averted by counter-measures taken in good time.

All this, in addition to improving care, frees the nurse from purely technical repetitive procedures and allows her to concentrate on those tasks, both manipulative and comforting, which only *she* can do. It is also probable that the standard of accuracy of measurements repeated very frequently, such as for instance quarter-hourly blood pressures, decays rapidly, with a tendency to err towards the normal. Under some special conditions, such as transplantation surgery or burns, it may be desirable, in the interests of preventing infection, to minimize the contact with the patient. Here monitoring provides the additional facility of making a measurement available at a distance. In coronary care also, it may be desirable to minimize activity around the patient, particularly if this is of an apprehension-generating kind.

What can be measured now, and how relevant is it?

Temperature:

Temperature can be measured at any site of the body surface, any accessible orifice and (by use of the "radio pill technique") in the gastro-intestinal tract, or at any location where surgical implantation is possible.

Because of the considerable temperature gradients which are possible within the body, the measurement at a particular site may be poorly related to core temperature. A patient, who because of extensive burns may be nursed without any covering in a ward where the temperature is not closely controlled, may exhibit a rectal temperature which is largely at the mercy of the temperature of the venous return from the legs. An ear temperature, unless the ear is well insulated, may be no better than a skin temperature.

Skin temperatures under certain conditions, can give valuable information about the circulation under the site of application. There is in fact, no such thing as a "patient's temperature", and the relevance of a measurement will depend on an understanding of the effect of the environment and the circulation at the measuring site.

Heart rate:

Heart rate may be measured from:

a. the EKG:

b. peripheral pulse transducer, which may either be of a displacement sensitive type, or use light reflection from sub-dermal layers. This tends to be less reliable than the EKG, it is subject to artefact, but gives a positive indication that a circulation is being maintained.

c. When an intra-arterial catheter is in place, the arterial waveform will give a measure of heart rate.

d. In theory at least, impedance plethysmography of any limb or digit can also be used.

In practice only methods a) and c) work reliably on any patient. Artefacts on b) tend to be due to movement, producing extra beats, so that a cessation of the signal might be taken as a warning that an adequate circulation is not being maintained.

In a recent trial under coronary care conditions, only 34 per cent of the crises were picked up by high or low limit alarms operating on heart rate alone. The situation might be improved by triggering alarms from a time × heart rate product so that for instance, a tachycardia of 150 beats/min does not itself trigger the alarm, but only does so after it has continued for some minutes—the required product perhaps decreasing with increasing heart rate.

The EKG waveform

In all intensive and coronary care wards, the oscilloscope display of the EKG waveform is the one piece of equipment which one finds almost always switched on; and about the usefulness of which there is least disagreement. Nurses can be taught to recognize arrythmias and to become competent at this, even when only using peripheral vision. However, if one considers that the average patient will produce of the order of 100,000 EKG complexes per 24 hours, it becomes obvious that only a small proportion of these are ever actually seen by the staff, and rare but important arrythmias may remain undetected until they occur with dangerous frequency. There is, therefore, a need for on-line automatic arrythmia detection. Various approaches, using both analog and digital computing techniques are making their appearance as commercial instruments; these I believe are capable of making a real contribution to patient care, though they are comparatively expensive.

Blood pressure

Arm cuff methods: A number of automatic cuff inflation instruments are available which, with one exception, use a microphone-like detector to detect the Korotkov sounds, and on the basis of the presence or character of the sounds make decisions about the level of the systolic and diastolic pressure. The exceptions are the Godart and Philips system which define the diastolic point by the abolition of a time delay between the arrival of the pulse wave at two points on the arm one above the other. All these instruments are to a greater or lesser degree subject to movement artefact. Few, if any, are reliable at systolic pressures below 75 mm. Hg., and most of these microphone types are critically dependent on the positioning of the microphone.

Finger cuff method: To measure systolic pressure only, finger cuffs are used, allied to pneumatic or piezo-electric detectors. The movement artefact problem is severe, and in any but a normal patient, the validity of the measurement is doubtful.

Intra-arterial methods: Catheter techniques used in conjunction with low displacement diaphragm-type transducers are well developed. Arterial pressure waveforms are displayed on oscilloscopes, and circuits have been devised for peak and trough picking to derive values of systolic and diastolic pressure. Contrary to usual practice, I do not believe that with good strain-gauge transducers and stable transistor operational amplifiers, more than very occasional calibration or zeroing is required (one in 24 hours). Automatic catheter flushing methods using miniature pumps have been developed which are not only safe, but make it possible to maintain catheters without any attention for several days. Catheter-tip transducers are available, though these tend to be fragile and very expensive.

The ethics of introducing a small catheter into a minor artery such as the radial artery tend to be regarded very differently in different hospitals. There is little evidence that this procedure when carried out with reasonable competence, and using a small catheter of suitable material, carries any appreciable risk. The statistics of arterial puncture appear to have been grossly distorted by the inclusion of cardiac catherisations and angiograms under the same heading as minor puncture carried out for the purpose of blood pressure measurement and blood sampling. Once a catheter is available it becomes a highway for hard physiological information. Apart from providing means for monitoring systolic and diastolic blood pressure, pulse rate and pulse amplitude, it should soon be possible to incorporate an oxygen partial pressure sensor, and maybe a thermistor in a disposable catheter.

One can thus look forward to the day when a patient's only connection to a monitoring system might be by one thin catheter and an EKG lead, accompanied by a marked improvement in the quality of the data.

Even the recognition and classification of arrythmias might be simplified by correlating the amplitude of the intra-arterial pulse wave with the EKG, rather than working entirely on the electrical signal.

Central (or any other venous) pressure

The equipment is similar to that used for arterial pressure. Because of the increased sensitivity required, zero stability presents a greater problem. From the operational point of view, however, movement of the patient relative to the gauge is a much more serious problem, because comparatively trivial changes in relative level can give rise to very misleading results. A special catheter is being developed which will place the reference pressure for a differential gauge at the same point within the patient, as the open end of the measuring lumen, to abolish this difficulty. Again, automatic flushing and debubbling make maintenance easier. Without automatic compensation, venous pressure cannot be *monitored* at all: it can at best be *measured* at intervals.

Respiration rate

This is a surprisingly difficult measurement, if patient acceptability is to be considered. Thermistors attached to the nose or mouth are one standard method, not well tolerated in conscious patients, though they work well in tracheotomy tubes. Impedance plethysmography of the chest is more acceptable, and by individural calibration could give some indication of tidal volume. If a central venous catheter is available, a respiration signal can be extracted from the pressure waveform.

A method which measures changes in front to back diameter of the rib cage, using the change in coupling between two very flat coils taped to the chest and back respectively, holds out some promise in being an alternative in impedance plethysmography.

If continuous pO_2 measurement on arterial blood becomes a routine procedure, then the measurement of respiration rate may decrease in importance.

Analysis of expired air

Physical gas analysers for oxygen and carbon dioxide are now available which are of sufficiently long-term stability to make continuous monitoring possible.

Urine flow

If a catheter has been inserted, either urine flow (by drop counting) or cumulative volume (by weight) can be recorded.

Haematocrit

Using conductivity methods, this can be measured on very small volumes of blood, and could therefore become a continuously available procedure, useful for instance, in the management of extensive burns.

Haemoglobin oxygen saturation

A catheter carrying two light guide bundles has been developed by the American Optical Company. This, with suitable external equipment, makes it possible to measure the haemoglobin oxygen saturation intra-vascularly by the "two wavelength" method. This equipment is, at present, very expensive, but would appear to have considerable clinical application—particularly if the catheter could be made sufficiently small to be manipulated into the venous outflow of individual organs.

Cardiac output

Both dye dilution and thermal dilution methods hold out some promise for comparatively frequent automatic cardiac output measurements. For the former. the special catheter mentioned above can be used instead of a cuvette.

Equipment layout

At one time discussion raged about the relative merits of central and bedside monitoring. This really only refers to the location at which the data is displayed, and becomes almost irrelevant in terms of equipment design if the display and alarm devices are produced as separate parts only requiring cable connection.

There is no doubt that indicators must be available at the bedside, otherwise the effect of procedures carried out on the patient cannot be directly evaluated. Whether they should also be available somewhere else, such as at a nursing station, depends on the staffing strategy and the geographical layout. If all beds can be seen from the nursing station, it is usually not difficult to arrange the displays so that these also can be seen. The solution does appear to be advantageous because it encourages even the single nurse to look at the patients, rather than "drive the ward from a cockpit". If this is not possible, then repeat displays will have to be provided in the nursing station together with alarms. (Figure 3).

What appears to be a poor solution is to provide the nursing station only with alarms, because at such times as the ward is monitored from that position it removes the main function of the system in providing trend information.

What the planner must always remember is that a monitoring installation is essentially a communication system between the patient and the staff responsible for his care. However clever the equipment is technically in terms of zero stability, accuracy and linearity, if it fails to communicate at the vital time it has failed utterly.

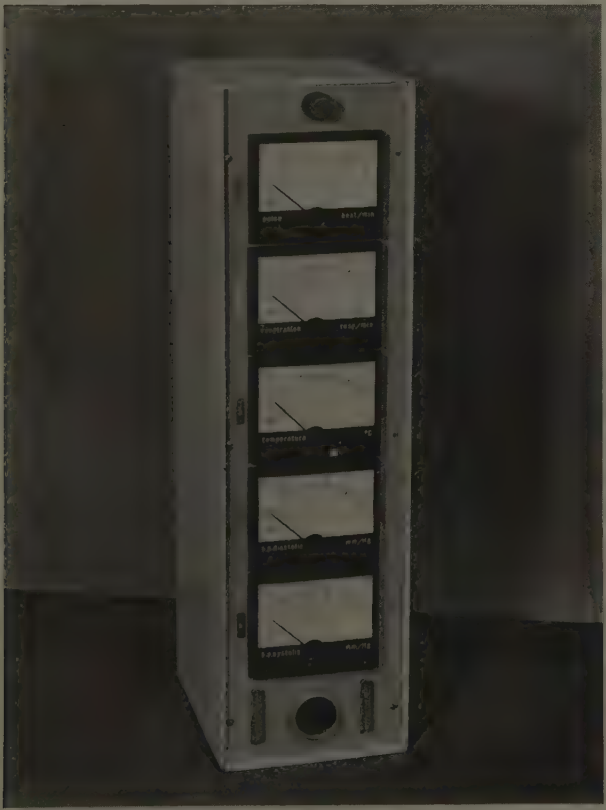


Fig. 3. A cheaper form of indication than that shown in Figure 2. Alarm conditions sensed at the bedside trigger a generally visible and audible alarm on this unit. The meters plug in and can be changed when the function of a channel changes.

Staff-equipment interaction

There is little doubt that in the vast majority of hospitals the technological climate, or perhaps better, the level of technological culture is insufficient to accept sophisticated monitoring equipment. This means that, in parallel with the introduction of the equipment, there must begin a deliberate and sustained programme of education for staff at all levels to familiarise them with the system and to point out its limitations.

Three types of training are required:

1. An appreciation on the part of the nursing and medical staff of what represents a significant difference between a value indicated by the monitor, and a value elicited by a traditional method. A nurse unless instructed otherwise, will assume that her method is error free, and that any deviation from this makes another method "wrong".
   It is also useful to stimulate amongst the staff an awareness of the amount by which a given measurement has to change before any clinical action is taken. The MCCCCM for "Minimum Change Capable of Changing a Clinician's Mind" (dubbed the 1600 effect) is a useful concept in determining the degree of accuracy for which it is reasonable to strive. An accuracy, resolution and stability of one half or one third of the MCCCCM should usually suffice.
2. Adequate training in the operation of the equipment. Until recently at least in Britian, there was a glaring disparity between the two-year training course for staff who were expected to operate simple X-ray equipment, and the few hours of demonstration and indifferent instruction books, which was all the nursing staff who were concerned with monitoring were likely to have. If one adds to that the fact that in many hospitals there is a rapid turnover of nurses in the intensive care ward, it is not suprising that often equipment is badly used or not used at all. Two things have to happen. First, the basic training of nurses must involve some familiarisation with the operation of technological devices, so as to remove the fear of the unknown when these are encountered. Secondly, the manufacturer must accept the responsibility for a much more rigorous system of staff training on installation, and for periodic refresher courses for nurses and technicians throughout the life of the equipment. This will, of course, cost money, but the user will have to accept that this represents part of the overhead costs of running an intensive or coronary care facility. In the USA there are now a few manufacturers who offer a package deal which does include planning, training and maintenance services, a development which is wholly salutary.
3. First line service by ward staff. Sooner or later equipment will fail, and it is at this point that staff attitude and training receive their most severe test. It is absolutely essential that as far as possible ordinary ward staff should be capable of recognising a fault and clearing it themselves. Once this state of affairs has been achieved, the system suddenly changes from a monster which can neither be controlled nor trusted, into a true servant of the staff. This means that the equipment must not only be modular, but that simple diagnostic devices

and procedures must be available which allow the defective module to be identified very quickly, so that it can be replaced from a stock of spares.

The penalty that must be paid for such an arrangement is an appreciable investment in spares, which might amount to 20 per cent of a small installation, and 10 per cent of a larger one. On the other hand technician support need not be so readily available, and service on the part of the manufacturer would in most cases be reduced to the repair of the defective module on his own premises. This should reflect itself in a reduction in running and first costs.

Figure 4 shows the bedside unit and Figure 5 shows the display system of a monitoring system which has been designed with this "Service without the Serviceman" concept in mind. Each of the signal conditioners (or normalisers) in the bedside unit is a complete module which can be removed without tools. Similarly each channel of the indicators is self-contained, can be removed, and only differs from its fellows by its scale plate.

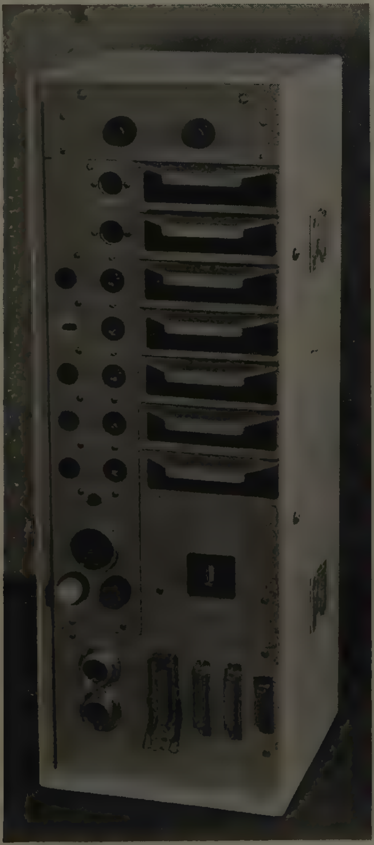


Fig. 4. The bedside unit. Each signal conditioning module can be removed and replaced without tools. The socket furthest on the bottom right takes the lead from one of the test instruments. (See Figure 7).

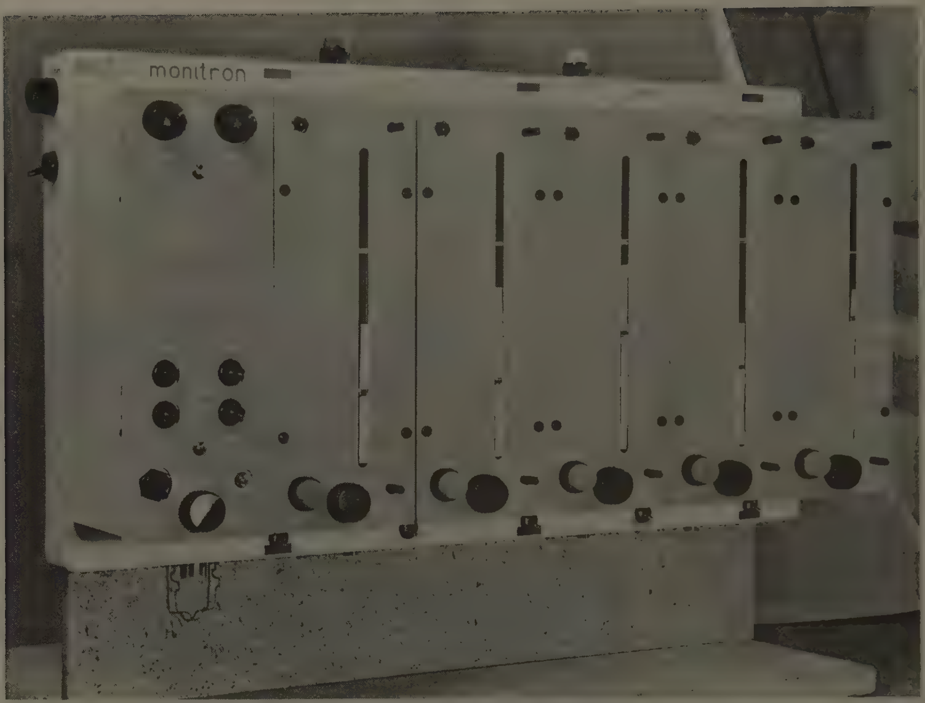


Fig. 5. An indicator bank. The individual indicators are electrically identical and only differ from each other by an easily removable scale plate.

The nurse is supplied with test programme charts like that shown in Figure 6, and three very simple test instruments as shown in Figure 7. When a fault

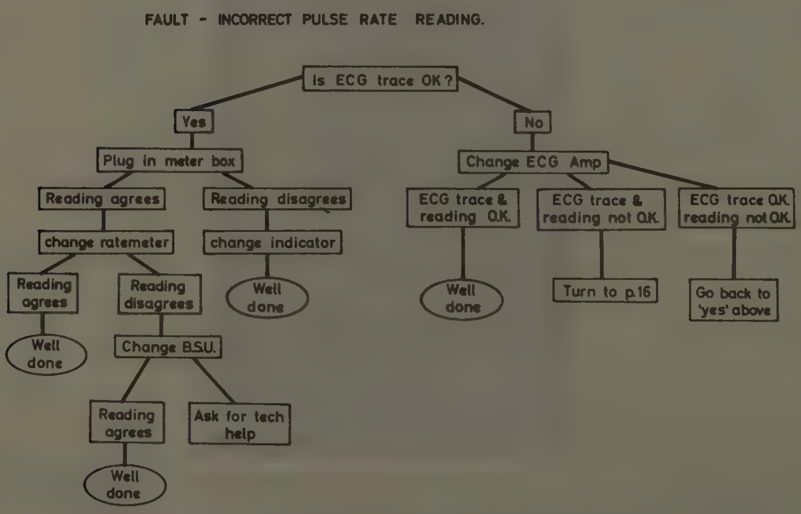

FAULT - INCORRECT PULSE RATE READING.
Is ECG trace OK ?
Yes
No
Plug in meter box
Change ECG Amp
Reading agrees
Reading disagrees
ECG trace & reading O.K.
ECG trace & reading not O.K.
ECG trace O.K. reading not O.K.
change ratemeter
change indicator
Reading agrees
Reading disagrees
Well done
Well done
Turn to p.16
Go back to 'yes' above
Well done
Change B.S.U.
Reading agrees
Ask for tech help
Well done

Fig. 6. A test programme chart.

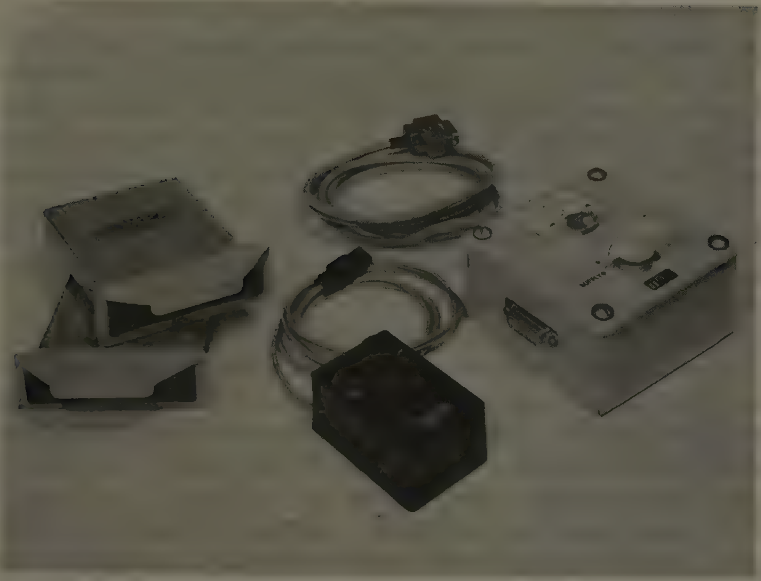


Fig. 7. The three types of test instrument required for the "Service without the Serviceman" concept. Furthest on the right is the meter unit, which can check the relationship between the output of signal conditioners and the indicators. Middle, simple signal source for testing patient leads, recorders, and oscilloscopes used in conjunction with diagnostic modules (right).

arises, or if she distrusts a particular measurement she turns to the appropriate chart and follows the directions. It is rare that the fault cannot be isolated after three or four branchings of the programme.

Care must be taken in the original design to avoid as far as possible faults in one module, leading to consequential faults in others. If this is not done it is possible for a situation to arise when a faulty module is quite correctly identified, but in which spares can be blown up *ad infinitum* as soon as they are inserted.

Technically quite unskilled people can operate the scheme described above, and I believe that the confidence which it breeds can be a major factor in rendering equipment acceptable and ensuring its full utilisation.

Development of monitoring technology

The present generation of equipment is basically restricted to the so-called vital signs and only performs minimal data processing. The next generation will provide facilities for continuous blood-gas and blood pH analysis and will make use of more sophisticated data processing techniques. Methods of display may be improved to give a more integrated picture of the overall physiological state of the patient.

In the third generation, I would predict the inclusion of some real online chemistry to bridge the gap between those substances which can be detected by electrode systems and those where the rate of change is sufficiently low to make the sending of a specimen to the laboratory a satisfactory procedure.

Phasic signal will be examined for long-term patterns by on-line digital computers, which will also have access to other data about the patient, and may make it possible to issue warning of potential deterioration at an earlier stage.

Monitoring systems will also begin to be used for diagnostic purposes, by examining the repercussions following the administration of a provocative procedure to the patient, thus measuring the ability of a given physiological system to maintain or regain homeostasis. This ought to give more information about the intactness of a system than the mere determination of the relationship of static measurements to average normality.

Contrary to popular opinion I believe that the vast majority of the information which can reasonably be required to monitor a patient in an intensive care situation, can be obtained using known and proven technology. What is much less well understood is the way in which complex equipment can be made to operate satisfactorily in an environment which is technically naive, and where powerful motivational forces may exist, which oppose the introduction of devices which on first sight put human vocation and skill at a discount. If one adds to this the loss of face which occurs every time when outside help has to be called to repair even the most trivial, but no less mysterious fault, it is not surprising that even good technology has not been enough to extract anything like the full potential benefit from monitoring equipment.

SOME PHYSICAL METHODS OF SKELETAL EVALUATION

John R. Cameron
John M. Jurist
Richard B. Mazess

Introduction

Changes in the musculoskeletal system occur in hypodynamic and hypogravic conditions; these changes, which occur in both bed-bound patients and astronauts, may lead to medical problems. This report summarizes some of the major changes, and reviews some newer physical methods useful in their measurement.

Disuse atrophy is a well-known clinical condition affecting both muscle and bone following denervation, prolonged bed rest, or immobilization. For many years investigators have speculated that weightlessness would have an adverse effect since the forces acting on the musculoskeletal system would be reduced, particularly in weight-bearing bones and postural muscles. Of course, only recently have studies in weightless conditions been possible. The limited mobility of astronauts suggests that marked alterations may occur in prolonged flight due to the "disuse" associated with both weightlessness and relative inactivity.

Bone loss poses two major hazards. First, ectopic mineral deposits, particularly in the kidney, may occur if bone loss is rapid or prolonged. Second, severe bone loss can compromise the structural integrity of the skeleton. Sensitive measurements of the skeleton are necessitated by these hazards in order to assess the nature and extent of changes in exposed individuals, and also to evaluate preventive and curative actions.

Changes of Skeletal Status

Various medical conditions have been examined, and several experimental techniques utilized, in order to determine the characteristics of bone changes with inactivity. Information has been derived from the study of myopathies, tenotomies, denervation syndromes, prolonged bed rest, fracture healing, as well as partial and total immobilization. In addition weightlessness has been simulated by water or oil immersion for relatively short time periods. Recent reports by other investigators (4, 7, 14, 17, 18, 23, 24, 52) have reviewed the major studies. There is an almost uniform indication that hypodynamic conditions, of whatever sort, in both animals and man result in hypercalciuria, negative calcium balance, and bone loss. The extent to which these findings may be ascribed to reduced generalized muscle activity, reduced weight-bearing stress, or altered neural, humoral, or blood flow patterns is still unclear.

Similar results have been obtained in space flight experiments. Mack and coworkers (26, 28, 29) used radiographic photodensitometry pre- and post-flight to evaluate bone mineral status. Average losses of 9, 12, and 14% were reported using os calcis radiographs of the Gemini IV, V, and VII crews. Losses in the metacarpals, oddly enough, were greater than those in the os calcis. The reported losses were not proportional to flight duration (4, 8, and 14 days), and much of the variation between flights was attributed to differences of calcium intake. Estimated dietary calcium intakes for the Gemini flights were about 710, 350, and 930 mg/day. Bedrest studies made by these same investigators supported the role of dietary calcium in ameliorating bone loss. However, several other studies have indicated that dietary influences on calcium mobilization in recumbency are minimal (5). Calcium and nitrogen balance studies during the Gemini VII mission provided limited information suggesting negative balance, particularly during the second week of the 14 day mission (53).

Average bone density losses have been reported for the three crew members of Apollo VII and of Apollo VIII (3); the average losses in the two flights were 0.8 and 4% of the os calcis, 1.2 and 10.4% for the distal radius, and 1.4 and 11.7% for the distal ulna. The individual bone changes ranged from +2 to −16%, and again the losses did not correspond to flight duration (260 hours for Apollo VII and 147 hours for Apollo VIII).

The extent of reported bone loss in the Gemini and Apollo crews was many times that expected for a comparable period of total immobilization. Further, the pattern of atrophy at different anatomical locations was variegated, and the losses in general did not correlate with duration of flight. It seems clear that undetermined factors are affecting these bone measurements, and it is possible that the radiographic method used was not sufficiently reliable.

The bone atrophy associated with hypodynamic conditions can reduce skeletal strength if continued for long periods. This is a major problem for older individuals who are subject to the usual loss of bone with age. Bed rest appears to induce a negative calcium balance of approximately 200-300 mg/day, and total immobility, as in poliomyelitis, may double this loss (52). Under severe hypodynamic conditions, the loss might amount to 2-4 g/week, and at least one month would be required for depletion of 1% of the body calcium store. Although atrophy is slow, continued loss would predispose toward fractures. The danger for astronauts could occur during re-entry deceleration. However, if the bone loss is irreversible, it may lead to problems when superimposed on the usual aging loss. This mobilization of bone raises the concomitant problem of abnormal mineral deposition. Such deposits, particularly urinary calculi, commonly occur in accelerated bone resorption syndromes, and calculi often have been noted in recumbency and inactivity.

The implications of bone atrophy in space flight seem generally adverse; the extent of bone loss predicted from hypodynamic conditions should not exceed a few percent per month, even if loss rates during space flight are several times more severe than during total immobilization. Such bone loss, if uniform, need not affect the structural integrity of the skeleton although it might stimulate extraskeletal mineralization. There are contrasting opinions on the risk and severity of possible kidney stone formation in space flight (11, 38); dehydration, metabolic alkalosis, and incomplete emptying of the urinary calyces may exacerbate this problem.

Methods of Measuring Skeletal Status In Vivo

Currently there are several practical methods for measurement of skeletal status which do not require bone biopsy or blood samples. These methods include direct photon absorptiometry, radiographic photodensitometry, and measurement of bone vibratory properties.

Measurement of the small rates of bone loss in hypodynamic and hypogravic conditions necessitates methods which are precise, accurate, and sensitive; these methods must also demonstrate long-term reliability. Further, there are changes of soft tissue composition under these conditions which may affect bone measurements. For example, shifts of body fluids occur with recumbency, and atrophy of muscle accompanies disuse (6, 41). In space flight, these problems would be compounded by the marked water loss seen in all flights to date (3). Alterations in the amount of soft tissue surrounding a bone, and changes in the fat, water, and muscle composition of that tissue, could affect measurements of bone mineral content determined by radiation absorption methods, and also might modify the measured vibratory properties of bone.

Direct Photon Absorptiometry

The photon absorptiometric method developed at the University of Wisconsin over the past 10 years uses the attenuation of a monoenergetic beam of low energy (20-100 kev) photons as a direct measure of bone mineral content (8, 9, 46). The basic principles of this method are illustrated in Figure 1. The bone mineral mass (M_B) at any point in the photon beam is proportional to ln (I^*_0/I) when the bone is embedded in a uniform thickness of soft tissue. The intensity of the beam is measured through the bone (I) and through the adjacent soft tissue (I^*_0). The bone mineral content can be determined at a single point, in a single path across a bone, or even over an entire anatomic area if serial scans are made.

A low energy radionuclide is used as a source of monoenergetic photons: ^{125}I (27.4 kev), ^{210}Pb (46.7 kev), and ^{241}Am (59.6 kev) have been used. Beam intensity is measured with a scinitillation detector-pulse height analyzer system. Scattered radiation is essentially eliminated by using a narrow, well-collimated beam and detector. We usually scan across a long bone, typically the radius, to deter-

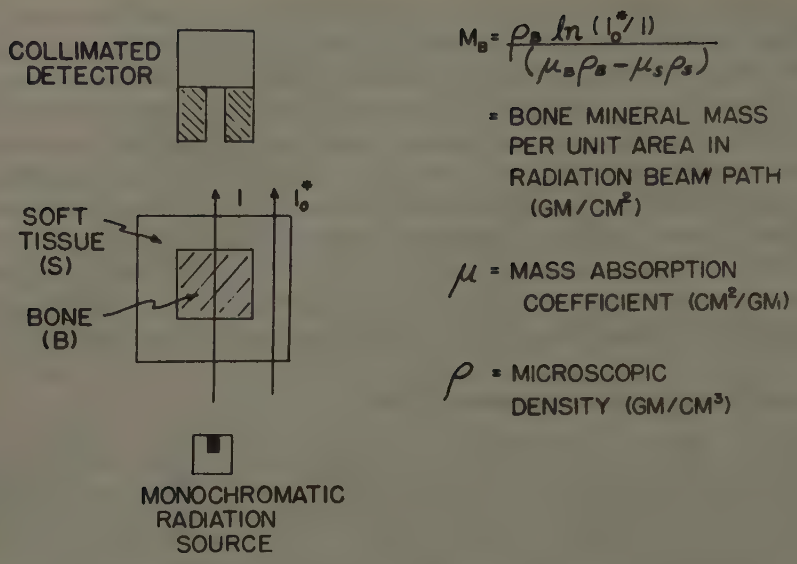

COLLIMATED DETECTOR
SOFT TISSUE (S)
BONE (B)
I
I_0^*
MONOCHROMATIC RADIATION SOURCE
$M_B = \frac{\rho_B \ln(I_0^*/I)}{(\mu_B\rho_B - \mu_S\rho_S)}$
= BONE MINERAL MASS PER UNIT AREA IN RADIATION BEAM PATH (GM/CM^2)
μ = MASS ABSORPTION COEFFICIENT (CM^2/GM)
ρ = MICROSCOPIC DENSITY (GM/CM^3)

Fig. 1. Basic principles used in measuring bone mineral with a monoenergetic photon source. The bone mineral mass per unit area in the beam path (M_B) is given by the equation. The measured quantities are I_0^*, the beam intensity through the soft tissue adjacent to the bone, and I, the beam intensity through the bone.

mine the mineral content of a given cross section. In such a case, the source and detector are mechanically linked and pass across the bone at a constant speed (Figure 2). The limb containing the bone of interest is surrounded by a soft tissue equivalent material in order to simulate uniform tissue thickness. We found either water or Super Stuff to work well. Super Stuff is approximately 95% water and is readily molded; it is available from Wham-O Corporation, San Gabriel, California or at local toy stores.

The measured beam intensity depends on the absorption characteristics of the material in the scan path. It is necessary to integrate the change of beam intensity across the scan path in order to determine the mineral mass in the path. This is usually done by recording measurements at fixed intervals across the bone with a scaler-timer, and then using computer analysis. We have also developed a direct readout unit which uses analog techniques to measure the mineral content in the scan path as well as its width. These values are presented immediately after completion of a scan; the unit is easily miniaturized and allows convenient monitoring of mineral content in patients, surveys, or in space flight.

The accuracy of the absorptiometric method has been demonstrated on ashed bone sections; section weight was predicted within about 3% over a wide range of values (10, 36). Varying the thickness and composition (muscle vs. fat) of surrounding tissue cover has little effect on the measurement (46). The method is very reproducible over periods of many months. Typical measurement precision is 1-2%, but this can be improved by scanning several times at a single location to

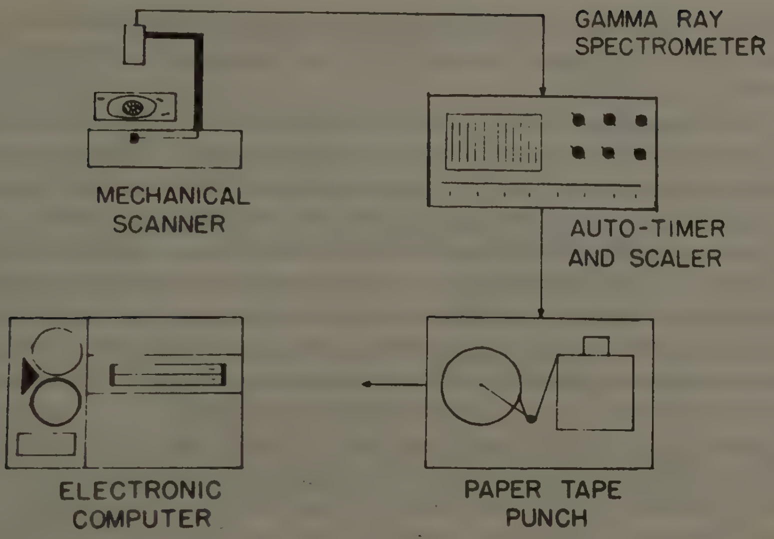


Fig. 2. System for measuring bone mineral content. Our laboratory is testing a direct readout unit which would allow determination of bone mineral content immediately after completion of a scan.

obtain an average. We are developing new techniques, using casts of the area of interest, to insure long-term precision and permit detection of small changes of mineral content.

The absorptiometric method is simple and involves a minimal radiation dose (10 mrad/scan). It is now used in at least a dozen medical centers for clinical and normative skeletal evaluation; the method has been used to evaluate changes in bed rest (47) and immobilization (22).

The same equipment used for absorptiometric measurement of bone suffices to measure soft tissue composition (fat vs. muscle + water) as well. The composition of a two-component system can be determined by making absorptiometric determinations at two selected photon energies (10, 45). We are using ^{125}I and ^{241}Am for these determinations. The measurement of the relative fat content of phantoms is accurate within about 2%, and measurement precision is about 1%. This method could be used to monitor dehydration or edema, fat changes, and muscle atrophy. The photon absorption method could therefore be used to monitor both soft tissue and bone changes occurring in hypodynamic or hypogravic conditions.

Radiographic Photodensitometry

The basic principles underlying radiographic photodensitometry are the same as for direct photon absorptiometry (27, 39, 50). The absorption of radiation by bone mineral is used as a measure of the mineral mass. However, instead of a monoenergetic radionuclide source this method uses a conventional X-ray tube, and a radiographic film is used instead of a scintillation detector. A photodensitometer is used to measure the optical density of the radiographic bone image as an indication of the mineral mass. An extensive annotated bibliography on this method is available (16).

Several problems arise in using this method which are not encountered with direct photon absorptiometry. The radiation from the conventional tube has a continuous energy spectrum, and absorption coefficients for such a beam cannot be well defined. In addition the heterogeneous beam changes in quality (hardens) as lower energy radiation is preferentially absorbed during passage through soft tissue and bone. There is no hardening of a monoenergetic beam, and absorption coefficients are well defined. Furthermore, an undetermined and variable amount of scattered radiation from the surrounding soft tissues affects the radiographic bone image; this does not occur with narrow beam geometry. Also, film usually exhibits a nonlinear response to both radiation intensity and energy.

Precision of a photodensitometer scan is usually quite high and this may have given the deceptive impression that the method itself is highly precise. Variations in film, film development, and film exposure conditions may contribute to large systematic errors. Standard exposure is a critical problem since the energy spectrum of a particular X-ray tube is poorly reproducible, and tubes differ greatly. These errors are partially reduced by using a calibrated wedge as a standard on each film. Investigators using this technique have sometimes erroneously failed to encase the bone of interest in a uniform thickness of tissue equivalent material and to bury the calibration wedge in this material.

Accuracy of radiographic photodensitometry has not been completely evaluated. Workers using the Pennsylvania State University technique, largely developed by Mack and associates, have reported high accuracy on excised bones and on bones with slight tissue cover. However, on bones with substantial tissue cover the error of the method is 20-50% (2, 43). Mayer et al. (35) concluded that scattered radiation was a major source of error. When appropriate methods are used to measure and compensate for scattered radiation (12), the predictive error of the measurement is reduced to about 6% on phantoms with moderate tissue cover (Colbert and Mazess, unpublished observations).

The overall utility of this method may be limited as much by long-term reliability as by accuracy. Numerous investigators report the usual precision to be approximately 4-8% (1, 25, 31, 35, 37). Even larger systematic errors are possible unless great care is taken during film exposure and development (48).

Radiographic photodensitometry has been used in a variety of clinical and normative investigations. It has been used to assess bone changes during inactivity, bed rest, and space flight (14, 26, 28, 29, 30).

Measurement of the Speed of Sound in Bone

Measurement of bone vibratory properties provides a non-absorptiometric method of skeletal evaluation which determines the functional characteristics of bone rather than its mineral mass. Several investigators have related the breaking strength of bone to its mineral content, density, elasticity, size, and shape as well as the age and sex of the propositus (13, 32, 33, 34, 49). Since the speed of sound in bone is related to its elasticity and density (21), several laboratories are studying acoustical or vibratory methods of skeletal evaluation. Rich and co-workers (42) exploited the difference between the speed of sound propagation in bone and in soft tissue to build a scanner for measurement of bone quantity. This technique worked reasonably well for compact bone, but the attenuation of ultrasound in bone is so great, especially if any cancellous portions are present, that their technique was impractical. Smith and Keiper (44) measured the elasticity of excised bone segments by a vibratory technique. Hyatt and co-workers (15) obtain the speed of sound in excised bone by ultrasound propagation and relate this measurement to the density and elasticity of the bone sample (Abendschein and Hyatt, to be published). It is generally recognized that bone is somewhat inhomogeneous with respect to elasticity (51), but an average or effective value may be measured.

Our laboratory is developing methods for measurement of the speed of sound in the ulna, tibia, clavicle, mandible, pelvis, and calvarium. We are studying the following approaches:

> Timing of mechanical impulse or ultrasound propagation along the bone. If two accelerometers are strapped to a long bone, the speed of sound can be obtained by timing a short-duration mechanical impulse between the transducers. Alternatively, the speed of ultrasound propagation along the bone can be measured. However, the large attenuation of ultrasound in bone, mentioned previously, makes this approach difficult.
>
> Measurement of the phase shift per unit length of bone for fixed vibrational frequencies. If a long bone such as the ulna or tibia is excited at one end by a sinusoidal vibration of known frequency, the relative phase difference between the outputs of two accelerometers placed on the bone, divided by the distance between the transducers, is proportional to the speed of sound in the bone at that frequency. Since shear waves are excited in the bone, the dispersion of phase velocity with frequency may be determined. This gives some information on the distribution of bone tissue.
>
> Measurement of resonant frequency. The product of resonant frequency and length of a long bone is related to the speed of sound in the bone (19, 20). Apparatus used to measure ulnar resonant frequency *in vivo* is shown in Figure 3. The resonant frequency is obtained from a recording of the amplitude response as a function of frequency (Figure 4). On repeated resonant frequency determinations, there is a standard deviation of 2-4%; the variation can be reduced by averaging several measurements. Determination of ulnar resonant frequency has been used for clinical investigations with promising results.
>
> Incorporation of bone into oscillator circuit. This is an alternate method of measuring the resonant frequency of the bone. The output from an accelerometer which measures response of the bone is connected to the input of a power amplifier which powers the electromagnetic driver used

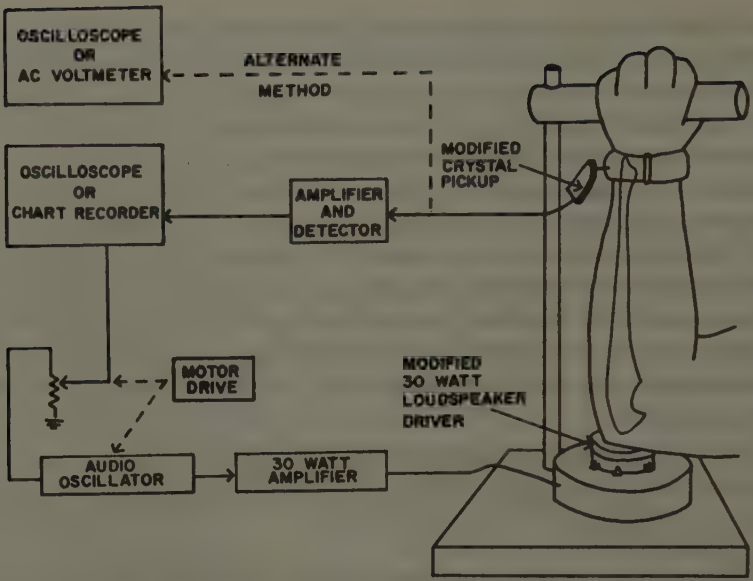

OSCILLOSCOPE
OR
AC VOLTMETER
ALTERNATE
METHOD
OSCILLOSCOPE
OR
CHART RECORDER
AMPLIFIER
AND
DETECTOR
MODIFIED
CRYSTAL
PICKUP
MOTOR
DRIVE
MODIFIED
30 WATT
LOUDSPEAKER
DRIVER
AUDIO
OSCILLATOR
30 WATT
AMPLIFIER

Fig. 3. Apparatus used for ulnar resonant frequency measurement. The driver is modified by bonding a small lucite piston to the driver diaphragm. The piston contacts the elbow of the subject. The ulnar response is detected by a small crystal accelerometer strapped to the wrist. With changes in the driver support, the apparatus may be used for measurement of tibial resonant frequency (Jurist, to be published).

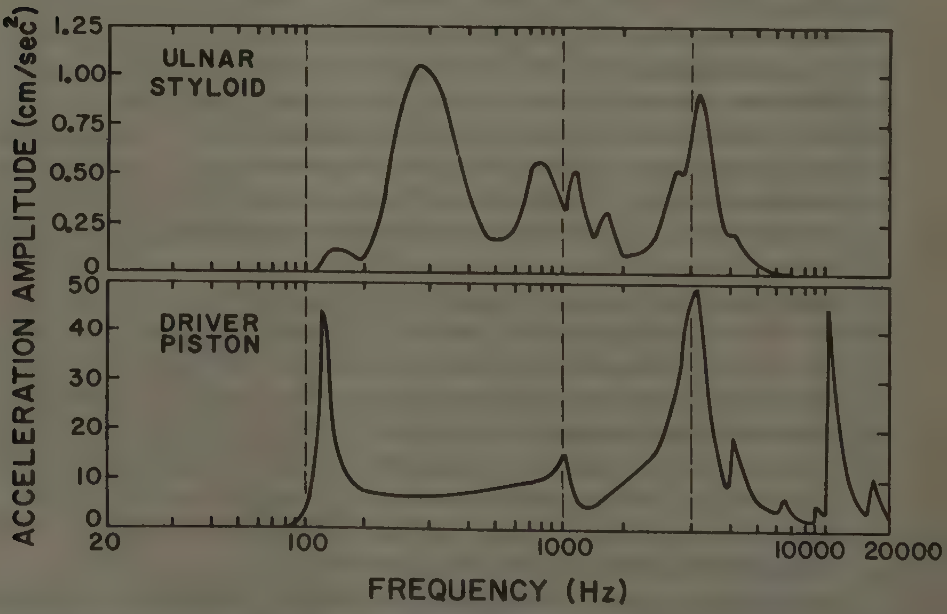


Fig. 4. Acceleration spectrum of the driver piston (bottom) and of the ulnar styloid (top) for a typical adult. Note the ulnar resonance at about 290 Hz. The resonances at about 130 Hz, 1000 Hz, and 3000 Hz are driver resonances. The resonance at about 10000 Hz is a resonance of the crystal accelerometer.

to excite the bone. The bone is, therefore, essentially the frequency controlling element of an oscillator circuit. The frequency of oscillation may be determined by using the oscillator output to drive a Schmitt trigger connected to a scaler.

Measurement of the transient response of the bone to application of a short-duration impulse. Analysis of the transient response by standard engineering methods allows determination of resonant frequency.

Discussion and Conclusions

Bone mineral losses in hypodynamic and hypogravic conditions probably are relatively slight compared to the total body mineral stores and major alterations would require long-term exposure. Findings of major atrophy in pre- and postflight radiographs are exceptional, but probably are attributable to the large experimental error of photodensitometry. The methods used for measuring bone changes under such conditions must be accurate, sensitive, precise, and reliable. A medical hazard is presented by ectopic mineral deposition. This hazard could be magnified by conditions in space flight. In all situations, the decision for dietary change, medication, ambulation, or other physical activity should be based on reliable measurements of bone status. Direct photon absorptiometry and measurement of bone elasticity afford convenient measurement of these critical parameters, and in addition photon absorptiometry can be used to monitor fluid and tissue changes. In contrast, radiographic photodensitometry is inaccurate on bones with moderate tissue cover; the possibility of large systematic errors makes the method very unreliable, even though short-term precision may be high. We have not discussed the possibility of utilizing mineral balance as an indication of skeletal status because of the great difficulties attendant to reliable, complete determinations. Balance studies are even difficult to use as research tools in metabolic wards. Although augmented calcium excretion may be a useful diagnostic tool, it need not reflect the extent of bone resorption. Neutron activation analysis may be another useful research tool (40), but the accuracy is unknown and the difficulties are great.

Newer physical methods are making nondestructive testing of the musculoskeletal system easier and more reliable. These methods will probably provide the most useful information when used in combination. The biomedical import of such information makes the exclusive use of any single method undesirable.

REFERENCES

1. Anderson J. B., J. Shimmins, and D. A. Smith., A new technique for measurement of metacarpal density. *Brit. J. Radiol.* **39**, 443-450, 1966.
2. Baker, P. T., H. Schraer, and R. G. Yalman., The accuracy of human bone composition determined from roentgenograms. *Photogram. Engr.* **25**, 445-460, 1959.
3. Berry, C. A., Preliminary clinical report of the medical aspects of Apollos VII and VIII. *Aerospace Med. 40*, 245-254, 1969.
4. Birge, S. J., and G. D. Whedon., Bone. in McCally, M. (ed.), *Hypodynamics and Hypogravics,* New York, Academic Press, 1968.

5. Birkhead, N. C., J. J. Blizzard, B. Issekutz, and K. Rodahl., Effect of exercise, standing, negative trunk and positive skeletal pressure on bed rest-induced orthostasis and hypercalciuria. AMRL-TR-66-6. Wright-Patterson AFB, Ohio, Aerospace Medical Research Laboratories, 1966.
6. Bourne, G. H., K. Nandy, and M. N. Golarz de Bourne., Muscle and the weightlessness state. in McCally, M. (ed.), *Hypodynamics and Hypogravics,* New York, Academic Press, 1968.
7. Browse, N. L., *The Physiology and Pathology of Bed Rest,* Springfield, Illinois, Charles C Thomas, 1965.
8. Cameron, J. R., and J. A. Sorenson., Measurement of bone mineral *in vivo. Science* **142,** 230-232, 1963.
9. Cameron, J. R., and J. A. Sorenson., A reliable measurement of bone mineral content *in vivo.* in *Proceedings of the Fifth European Symposium on Calcified Tissues,* Paris, 1969.
10. Cameron, J. R., R. B. Mazess, and J. A. Sorenson., Precision and accuracy of bone mineral determination by direct photon absorptiometry. *Inv. Radiol.* **3,** 141-150, 1968.
11. Cockett, A. T. K., C. C. Beehler, and J. E. Roberts., Hypodynamic urolithiasis: A potential hazard during prolonged weightlessness in space travel. *School of Aerospace Medicine Review* 2-62. Brooks AFB, Texas, 1961.
12. Colbert, C., J. J. Spruit, and L. R. Davila., Biophysical properties of bone: Determining mineral concentration from the X-ray image. *Trans. N.Y. Acad. Sci., Series II.* **30,** 271-290, 1967.
13. Currey, J. D., The mechanical consequences of variation in the mineral content of bone. *J. Biomech. 2,* 1-11, 1969.
14. Dick, J. M., Objective determinations of bone calcium levels. *Aerospace Med.* **37,** 136-139, 1966.
15. Floriani, L. P., N. T. Debevoise, and G. W. Hyatt., Mechanical properties of healing bone by the use of ultrasound. *Surg. Forum.* **18,** 468-470, 1967.
16. Garn, S. M., An annotated bibliography on bone densitometry. *Am. J. Clin. Nutr.* **10,** 59-67, 1962.
17. Hattner, R. S., and D. E. McMillan., Influence of weightlessness upon the skeleton: A review. *Aerospace Med.* **39,** 849-855, 1968.
18. Heaney, R. P., Radiocalcium metabolism in disuse osteoporosis in man. *Am. J. Med.* **33,** 188-200, 1962.
19. Jurist, J. M., Ulnar vibratory properties. in *Proceedings of the Conference on Progress in Methods of Bone Mineral Measurement,* Bethesda, USPHS-NIAMD, 1969.
20. Jurist, J. M., and W. A. Selle., Acoustical detection of osteoporosis. *The Physiologist.* **8,** 203, 1965.
21. Kinsler, L. E., and A. R. Frey., *Fundamentals of Acoustics,* New York, John Wiley & Sons, Inc., 1950.
22. Lundberg, B. J., and B. E. Nilsson., Osteopenia in the frozen shoulder. *Clin. Orthop.* **60,** 187-191, 1968.
23. Lynch, T. N., R. L. Jensen, P. M. Stevens, R. L. Johnson, and L. E. Lamb., Metabolic effects of prolonged bed rest: Their modification by simulated altitude. *Aerospace Med.* **38,** 10-20, 1967.
24. McCally, M., The pathophysiology of disuse and the problem of prolonged weightlessness: A review. AMRL-TDR-63-3. Wright-Patterson AFB, Ohio, Aerospace Medical Research Laboratories, 1963.
25. McFarland, W., Evaluation of bone density from roentgenograms. *Science* **119,** 810-811, 1954.
26. Mack, P. B., and P. A. LaChance., Effects of recumbency and space flight on bone recumbency. *Am. J. Clin. Nutr.* **20,** 1194-1205, 1967.
27. Mack, P. B., W. N. Brown, and H. D. Trapp., Quantitative evaluation of bone density. *Am. J. Roent.* **61,** 808-825, 1949.
28. Mack, P. B., P. A. LaChance, G. P. Vose, and F. B. Vogt., Review of medical findings of Gemini VII and related missions-bone demineralization. in *A Review of Medical Results of Gemini 7 and Related Flights,* Washington, D.C., NASA Space Medicine Directorate, 1966.
29. Mack, P. B., P. A. LaChance, G. P. Vose, and F. B. Vogt., Bone demineralization of foot and hand of Gemini-Titan IV, V, and VII astronauts during orbital flight. *Am. J. Roent.* **100,** 503-511, 1967.

30. Mack, P. B., R. A. Hoffman, and A. N. Al-Shawi., Physiologic and metabolic changes in *Macaca nemestrina* on two types of diets during restraint and non-restraint. II. Bone density changes. *Aerospace Med.* **39,** 698-704, 1968.
31. Mason, R. L., and C. Ruthven., Bone density measurements *in vivo:* Improvement of X-ray densitometry. *Science* **150,** 221-222, 1965.
32. Mather, B. S., Comparison of two formulae for *in vivo* prediction of strength of the femur. *Aerospace Med.* **38,** 1270-1272, 1967.
33. Mather, B. S., Correlations between strength and other properties of long bones. *J. Trauma* **7,** 633-638, 1967.
34. Mather, B. S., The effect of variation in specific gravity and ash content on the mechanical properties of human compact bone. *J. Biomech.* **1,** 207-210, 1968.
35. Mayer, E. H., H. G. Trostle, E. Ackerman, H. Schraer, and O. D. Sittler., A scintillation counter technique for the X-ray determination of bone mineral content. *Rad. Res.* **13,** 156-167, 1960.
36. Mazess, R. B., J. R. Cameron, R. O'Connor, and D. Knutzen., Accuracy of bone mineral measurement. *Science* **145,** 388-389, 1964.
37. Morgan, D. B., F. W. Spiers, C. N. Pulvertaft, and P. Fourman., The amount of bone in the metacarpal and the phalanx according to age and sex. *Clin. Radiol.* **18,** 101-108, 1967.
38. Nungesser, W. C., Factors influencing the renal regulation of calcium-implications of prolonged weightlessness. *School of Aerospace Medicine Review* 2-65. Brooks AFB, Texas, 1965.
39. Omnell, K. A., Quantitative roentgenologic studies on changes in mineral content of bone *in vivo. Acta Radiol. Suppl.* 148, 1957.
40. Palmer, H. E., W. B. Nelp, R. Murano, anr C. Rich., The feasibility of *in vivo* neutron activation analysis of total body calcium and other elements of body composition. *Phys. Med. Biol.* **13,** 269-279, 1968.
41. Piemme, T. E. Body fluid volume and renal relationships to gravity. in McCally, M. (ed.), *Hypodynamics and Hypogravics,* New York, Academic Press, 1968.
42. Rich, C., E. Klink, R. Smith, B. Graham, and P. Ivanovich., Sonic measurement of bone mass. in *Progress in Development of Methods in Bone Densitometry,* Washington, D.C., NASA SP-64, 1966.
43. Schraer, H., R. Schraer, H. G. Trostle, and A. D'Alfonso., The validity of measuring bone density from roentgenograms by means of a bone density computing apparatus. *Arch. Biochem. Biophys.* **83,** 486-500, 1959.
44. Smith, R. W., and D. A. Keiper., Dynamic measurement of viscoelastic properties of bone. *Am. J. Med. Elect.* **4,** 156-160, 1965.
45. Sorenson, J. A., and J. R. Cameron., Body composition determination by differential absorption of monochromatic X-rays. in *Proceedings of the Symposium on Low Energy X- and* Gamma Sources and Applications, Washington, D.C., AEC Report ORNL-IIC-5, 1964.
46. Sorenson, J. A., and J. R. Cameron., A reliable *in vivo* measurement of bone-mineral content. *J. Bone Jt. Surg.* **49A,** 481-497, 1967.
47. Vogel, J. M., Changes in bone mineral content of the os calcis induced by prolonged bedrest. *Fed. Proc.* **28,** 374, 1969.
48. Vose, G. P., Factors affecting the precision of radiographic densitometry of the lumbar spine and femoral neck. in *Progress in Development of Methods in Bone Densitometry,* Washington, D.C., NASA SP-64, 1966.
49. Vose, G. P., and A. L., Kubala., Bone strength-its relationship to X-ray determined ash content. *Hum. Biol.* **31,** 262-270, 1959.
50. Vose, G. P., S. A. Hoerster, and P. B. Mack., New technics for radiographic assessment of vertebral density. *Am. J. Med. Elect.* **3,** 181-188, 1964.
51. Walmsley, R., and J. W. Smith., Variations in bone structure and the value of Young's modulus. *J. Anat.* **91,** 603, 1957.
52. Whedon, G. D., Osteoporosis: Atrophy of disuse. in Rodahl, K., J. T. Nicholson, and E. M. Brown (eds.), *Bone as a Tissue,* New York, McGraw-Hill Book Co., 1960.
53. Whedon, G. D., L. Lutwak, and W. Newman., Calcium and nitrogen balance. in *A Review of Medical Results of Gemini 7 and Related Flights,* Washington, D.C., NASA Space Medicine Directorate, 1966.

AUTOMATED TECHNIQUES FOR THE BEDSIDE OF THE CRITICALLY ILL

M. H. Weil, H. Shubin, R. Martin,
J. H. Carrington and N. Palley

INTRODUCTION

During the past decade a number of specialized hospital units have been developed for monitoring and care of critically ill patients. These units, which include intensive care, coronary care, trauma, renal dialysis and post surgery recovery units, account for between four and eight per cent of the bed capacity in community hospitals.

Assessment of the circulatory and respiratory status of critically ill patients requires repetitive measurement of those variables which indicate the physiological integrity of the patient. Hemodynamic measurements quantitate arterial and venous pressure, blood flow (cardiac output) and blood volume. Electrical parameters such as the electrocardiogram and the electroencephalogram are recorded. The oxygen tension and saturation, carbon dioxide tension and pH are measured on samples of arterial blood as well as electrolyte and organic constituents of blood (potassium, sodium, chloride, urea, glucose and lactic acid). Continuing or repetitive assessment is required since the critically ill patient is likely to have rapid and often unpredictable changes in status (1).

For purposes of monitoring and competent clinical care, the organization and operation of critical care units require a major commitment of highly trained physicians, nurses and technicians. The human talents, though primary, must be supported by substantial physical resources for a relatively small number of patients. In view of the numerous diagnostic and therapeutic services that are now involved, major improvements in the efficiency of operation are mandatory if the availability of professional personnel and the full resources of supplies and equipment are to accrue for the benefit of patient care (2).

Clinicians are hampered not only by the complexity of the methods by which data is secured but also with the often overwhelming task of data analysis and interpretation for immediate use at the bedside. The availability of such information has been life saving. Present manual methods of data collection and logging impose great demands on the staff, particularly during night hours and weekends when staffing is reduced. This has encouraged physicians and paramedical personnel to foster the use of automated data processing methods to expedite both the acquisition and analysis of measurements at the bedside of the critically ill.

Hemodynamic measurements on patients

The routine use of intravascular catheters, electrodes and sampling devices has proven to be a technically formidable undertaking. Some of the difficulty is directly attributable to the limited engineering investment in clinical instrumentation. Available instrumentation is unduly complex and the operational difficulties that are encountered in its routine use reflect differences in dialogue between designer and user. Routine operation of commercially available amplifiers usually imposes a requirement for repetitive adjustment of 5 or more controls, and the accuracies of the measurements themselves are limited by inconsistencies in technical operation, the variability of calibration standards, and the formidable procedures involved in error checking. Physical restraints are often imposed by sensors and transducers, particularly by the cables with which they are connected to the remainder of the system. They compromise access to the patient by physicians and nurses. These problems are magnified in the critical care environment in which the technical demands for operation of monitoring equipment is construed as competitive with the professional needs for superior patient care.

With the resources provided by the interdisciplinary team of physician-physiologists, engineers, and biomathematicians of our Shock Research Unit, major improvements in the operational design of a system for measuring intravascular pressures, electrocardiographic signals, and body temperatures are achieved (Figure 1). Sensors and transducers are connected by short cables to an instrument island which was so constructed that access to the patient and mobility of the bed is retained. The transducers themselves are housed in movable receptacles so that their position in relationship to the body of the patient and the maintenance of their connection to the remainder of the system are optimized (Figure 2).

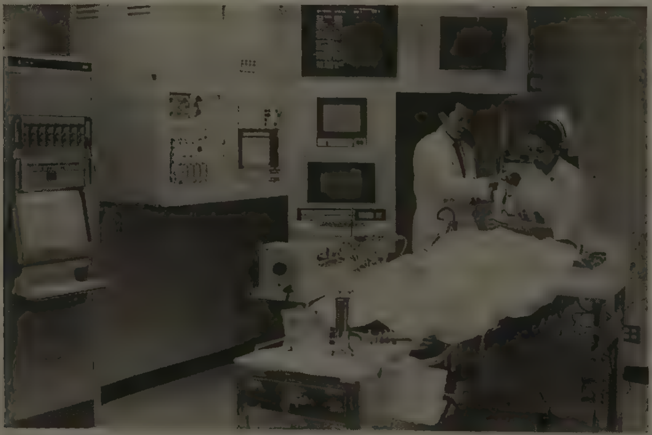


Fig. 1. Photograph of one of the two automated bed modules in the U. S. C. Shock Research Unit supported by a Sigma 5 digital Computer; digital and analog display of patient status is provided on one of four oscilloscopes. The computer, plotter and teletype devices for creation of the patient record and a high speed printer are located in a remote area.

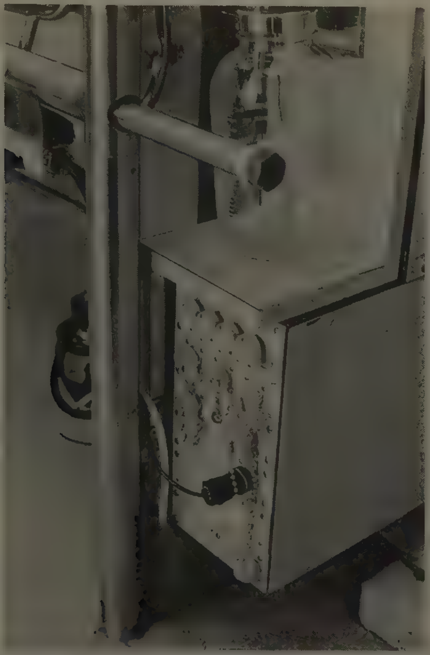


Fig. 2. Portion of a bedside island which facilitates connection of instrumentation to the patient and decreases physical restraint by cables and other connections.

Sizable electronic components are maintained in a remote position adjacent to but without impingement on the bedside area. All of the amplifiers were designed to have very low output impedances to minimize incompatibilities that would result from multiple inputs into a common data acquisition system. Full-scale values were standardized in all amplifiers and provisions were made to safeguard against short circuits and imperfect grounding.

With special reference to the blood pressure amplifiers, calibration adjustment is accomplished by two internal servo motors and servo loops. These serve to control valves which open the system to reference pressures which are applied to the transducers. Motor drive control relays are activated and the completion of the calibration procedure is sensed. By these techniques, a zero reference pressure is translated into a zero voltage output and a reference pressure of 200 mm Hg is translated, by servo adjustment, to a four volt output. Since the pressure source has an inconsistency of only 0.1 per cent, this simplified calibration procedure also provides notable improvement in accuracy of the pressure reference. The entire automated procedure is accomplished by the sequential depression of two switches, one for the zero reference signal and the other for the 200 mm Hg calibration signal. The amplifier is also provided with a test circuit which activates a light indicator in the event that the calibration procedure is not accomplished. The amplifiers are physically small and of modular configuration to allow clinical staff to conveniently exchange units in case of failure during the period of clinical use (Figure 3).

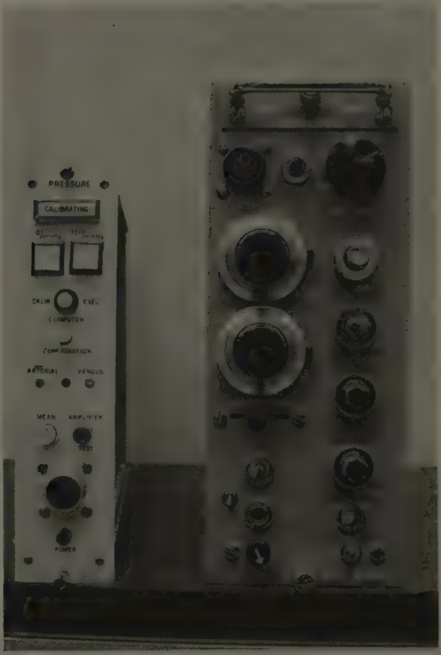


Fig. 3. Comparison between standard, commercially available blood pressure amplifier and the newly designed module which may be operated by the use of only two switches.

The electrocardiographic amplifier is completely isolated from ground and the patient circuit operates at a maximum of 6 volts. Again, only two controls are provided, a calibration switch and a lead selector. Comparable concepts are inherent in the design of amplifiers for measurement of temperature and for measurement of urine flow (Figure 4). Maximal ease of calibration, error checking methods, and compatibility to data acquisition equipment are assured.

In its present configuration, these components were designed to operate as part of a larger digital-computer controlled patient monitoring system. However, the components of the system and particularly the amplifier designs which have been developed quite apart from their relationship to the computer serve as examples of designs by which operational simplification and error checking is achieved. Such simplification and insurance of reliability is an important goal if physiological measurements are to be routinely available at the bedside of the critically ill.

Automation

Automated techniques, now bolstered by the availability of a computer, have become increasingly important in recent years as medicine seeks to fulfill the need for ever more comprehensive and technically complex services (3,4).

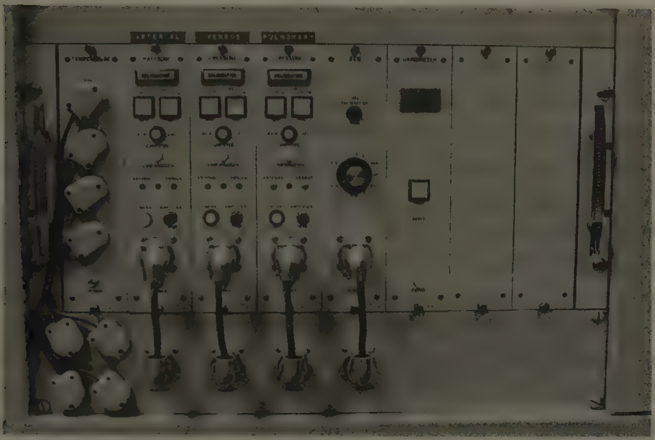


Fig. 4. The bank of compact, self calibrating amplifiers and controllers is shown. Signal conditioning devices prepare the "outputs" of the physiological sensors on patients to serve as "inputs" to clinical computer system.

Detailed analysis of physiological data by manual methods is inefficient, both for clinical research and for practical use at the bedside. A combination of sensors with analysis of their output on an on-line computer, offers a means of achieving a comprehensive inventory of patient condition for immediate use by the physician and the clinical investigator.

Automated measurement and collection of biological data in a clinical setting has been used in a number of areas. Automated analog devices have been developed for measuring blood pressure by the use of indirect auscultatory techniques. Analog computers have been developed to determine cardiac output by integration of the corrected area under the indicator dilution curve. Flow through large blood vessels has been measured by ultrasound and impedance techniques in patients during surgery. Analog devices making use of the electrocardiographic signal for measuring and reporting heart rate and detection of disturbances of the heart rhythm have assumed an important role in the monitoring of patients in critical care areas. Respiratory ratemeters detect pressure or impedance changes during chest movements. Automatic measurement of rectal and esophageal temperatures by the use of thermistors is common during anesthesia, particularly during cardiovascular surgery. Analog devices have been used for chemical determinations such as blood glucose, lactate, pH, carbon dioxide pressure (PCO_2), and oxygen pressure (PO_2). Telemetry techniques are employed to facilitate transmission of the signal to remote recorders. Both analog and digital computers have also assumed an increasingly important role for data control and efficient processing of data generated in clinical laboratories. The contribution that automation makes when applied to each of these techniques greatly bolsters the wider availability as well as the accuracy of these measurements (5).

Computer operation

An important development, for which the Shock Research Unit shares responsibility, is the use of a combination of sensing devices coordinated by a digital computer to develop an integrated picture of a patient's physiological condition on a time-related basis. Dynamic signals including pressure and pulse measurements are continuously presented on a bedside oscilloscope by signal conditioning hardware as shown in Figure 4. The signal conditioners amplify the signals and perform a preliminary analysis of wave forms. The output of these conditioners is attached to the multiplexer, then to the A/D converter interfaced to a computer channel. The computer must also be capable of real time analysis of these signals. As an example, the computer must be capable of recognizing individual components of the EKG signal (i.e., the QRS wave), find the pulse wave in the arterial pressure signal, calculate maximum, minimum, and mean arterial pressure, before the next heart beat occurs. Although the heart rate is normally about 80 per minute, rates in excess of 160 per minute are often encountered in the critically ill patient and these wide ranges of values are taken into account (3,6,7,8).

An on-line blood chemistry unit is being developed for use at the bedside. This will enable the computer to monitor PO_2, PCO_2, pH and electrolytes as well as blood volume measurements. The timing for this type of monitoring must take into consideration the mechanical delay between sampling and completion of the measurement. Measurements may be as simple as the sampling of the output of an electrometer connected to an electrode or as complex as spectral analysis.

Data reduction

The computer's ability to analyze signals at beat-to-beat rates is an asset during moments of crisis and during procedures such as drug administration. However, if such detailed information were accumulated continuously, the physician would be overwhelmed by a flood of data. A primary schedule of monitoring and reporting is needed which will insure that the patients are under constant surveillance, but the physician and computer are relieved of an overwhelming problem of data accumulation and management by the use of statistical limit setting techniques.

The monitoring schedule now available in our Unit allows for two modes of operation. The physician is able to request beat-by-beat measurements. The routine mode of monitoring, however, is based upon the ability to define or predict the amount and rate of change to be expected in any given variable. Heart rate and rhythm are monitored almost continuously because abrupt changes in these variables may be followed by rapid deterioration in patient status. However, skin and rectal temperature are monitored at 15 minute intervals because significant changes in these signals are generally of longer duration.

After the data are converted into physiological units, they are statistically analyzed to determine this significance. The change in a specific variable is related to expected changes in other measurements accumulated at the same time. If the variations are statistically not significant the data are not preserved. However, if there has been a significant change, this is logged in the computer records. The ward staff is immediately alerted and informed as to which variable has changed.

Effect on medical practice

We recognize unusual opportunity for the use of automation at the bedside to reduce the routine demands on the physician and particularly on the nurses, much of whose time is devoted to purely mechanical chores. Is there special advantage for the patient who is critically ill if measurements of temperature, blood pressure, and respiratory rate are accomplished by manual techniques rather than by automated sensors? Is the patient either psychologically or physically more appropriately managed if the intravenous infusion rate is adjusted and readjusted at frequent intervals by manual rather than by highly reliable mechanical means to deliver the volumes of fluid which have been ordered by the physician? Is it better to measure the output of urine with a graduated glass cylinder than to utilize a urinometer which provides a continuous and highly accurate indication of urine flow? Would the patient in fact lose identity because of the improvement in efficiency and reliability of such automated measurements? Experience over the past six years testifies to the contrary. Mechanization frees the physician and nurse to provide the very empathy and warmth in doses far exceeding those previously possible.

The clinician who has available to him a system at the bedside is fully cognizant that the machine is only a tool to assist the professional efforts of the physician and nursing staff and no more. If the efficiency of automation improves the scientific competence of the physician by extending the scope and accuracy of his judgment, and particularly if it frees him to attend more closely to the decision making and humanitarian aspects of care, the machine fully justifies its existence. It does so in the same way that the scalpel has an advantage over the kitchen knife and the measurement of temperature for documentation of fever compares favorably with even the most perceptive mother's palm (4). The computer based system is neither expected to nor called upon to make medical judgments. It is expected to—and does—rapidly provide the physician with reliable information. By freeing the physician from many routine tasks of patient monitoring it enables him to devote himself more fully to critical judgments in patient care.

The acceptance of and interest in a system of the type described by the medical faculty and house staff at our hospital have assured us that responsible physicians are not generally fearful of nor distracted by gadgets and computers. They rapidly assess its proper place as a useful accessory. They welcome its efficiency and particularly the speed and accuracy with which it provides them with data at critical times and reject its use if its advantages do not outweigh technical complexity for routine use.

Summary and conclusions

Current medical knowledge is such that remarkably effective treatment can be offered to patients who present with acute life-threatening injury or illness. Yet, the ordinarily overwhelming complexity of medical maneuvers at the bedside of such patients and the limitations of available professional staff and facilities preclude the full application and utilization of important medical accomplishments of proven value. Fortunately, the feasibility of the systems approach to bedside care and automation of repetitive tasks has already proven to be suitable and helpful. Even if the rapidly increasing costs for personnel were of little concern, the limited availability of technical personnel and the inordinately complex facilities for monitoring and management now used at the bedside seemingly leave no alternative to mechanization and automation. Most important of all, if competent and wide use of remarkable advances in technology are to accrue to patients at a time of life-determining need, generous provision will have to be made for automation.

BIBLIOGRAPHY

1. Weil, M. H., Computer automation of hospitals, Guest Editorial. *Journal of the Amer. Hosp. Assoc.* **40:** 71, September, 1966.
2. Weil, M. H., A new look at the critically ill. *Hosp. Prac.* **2:** 27-28, January, 1967.
3. Jensen, R. E., Shubin, H., Meagher, P. F. and Weil, M. H., On-line computer monitoring of the seriously ill patient. *Med. & Biol. Engr.* **4:** 265-272, 1966.
4. Weil, M. H., Shubin, H., and Rand, W. M., Experience with a digital computer for study and improved management of the critically ill. *JAMA* **198:** 147-152, November, 1966.
5. Warner, H. R., The role of computers in medical research. *JAMA* **196:** 944-949, June, 1966.
6. Wilbur, S. A. and Derrick, W. S., Physiological monitoring during anesthesia and surgery. Proceedings Fourth Natl. Biomedical Sciences Instrumentation Symposium **3:** 197-206, 1966.
7. Talbot, G. Douglas, The Cox coronary heart institute prototype for an integrated automated system. Proceedings Fourth Natl. Biomedical Sciences Instrumentation Symposium **3:** 161-169, 1966.
8. Warner, H. R., Gardner, R. M., Toronto, A. F., Computer-based monitoring of cardiovascular functions in post-operative patients. Circulation 37 (supple. 2): II-68-II-74, April, 1968.

HOLOGRAPHY AND ITS APPLICATION TO MEDICAL PHYSICS

Ernest Feleppa

INTRODUCTION

Two decades ago Dennis Gabor conceived a method of overcoming the problem of spherical aberration in electron microscopy (1,2). At that time he formulated the theory of this method, holography, but for a number of technical obstacles which have not yet been overcome, holography was not realized as a practical technique in electron microscopy. However, with the invention of the continuous, visible-light laser, the feasibility of Gabor's technique was quickly recognized (3-5). Physical scientists have developed holography so that, in its many ramifications, it now appears to be a potentially significant and valuable technique in a broad spectrum of research.

For example, it merits serious attention as a useful complement to the physical tools of medicine. Not only can holograms provide phase-faithful, three-dimensional images, but they can be used as complex, spatial filters to recognize patterns or to improve the resolution and contrast of noncoherent images.

IMAGING

A hologram is a recording of the interference pattern arising from the superposition of two coherent waves (1-5). In imaging applications, one of these waves is the wave scattered from the object of interest and the other is a simple spherical or plane reference wave (Fig. 1-a.). Upon illumination, at a later time, by a third coherent wave, the hologram will diffract the illuminating wave, forming two image-containing waves (Fig. 1-b.).

The usual holographic apparatus (Fig. 2) consists of a laser as a source of coherent light, a beam-splitting device to provide the object-illuminating and reference waves, the object, and a recording medium. With visible light, the hologram is usually recorded on a photographic emulsion. A number of lenses and mirrors might be incorporated in the apparatus to direct and shape the waves. At the recording plane, the reference wave, U_1, and object-scattered wave, U_2, combine to produce a resultant wave, $U_1 + U_2$. U_1 and U_2 are complex quantities describing the amplitude and phase of the waves as a function of spatial coordinates. The emulsion will record the time-averaging intensity of $U_1 + U_2$, or

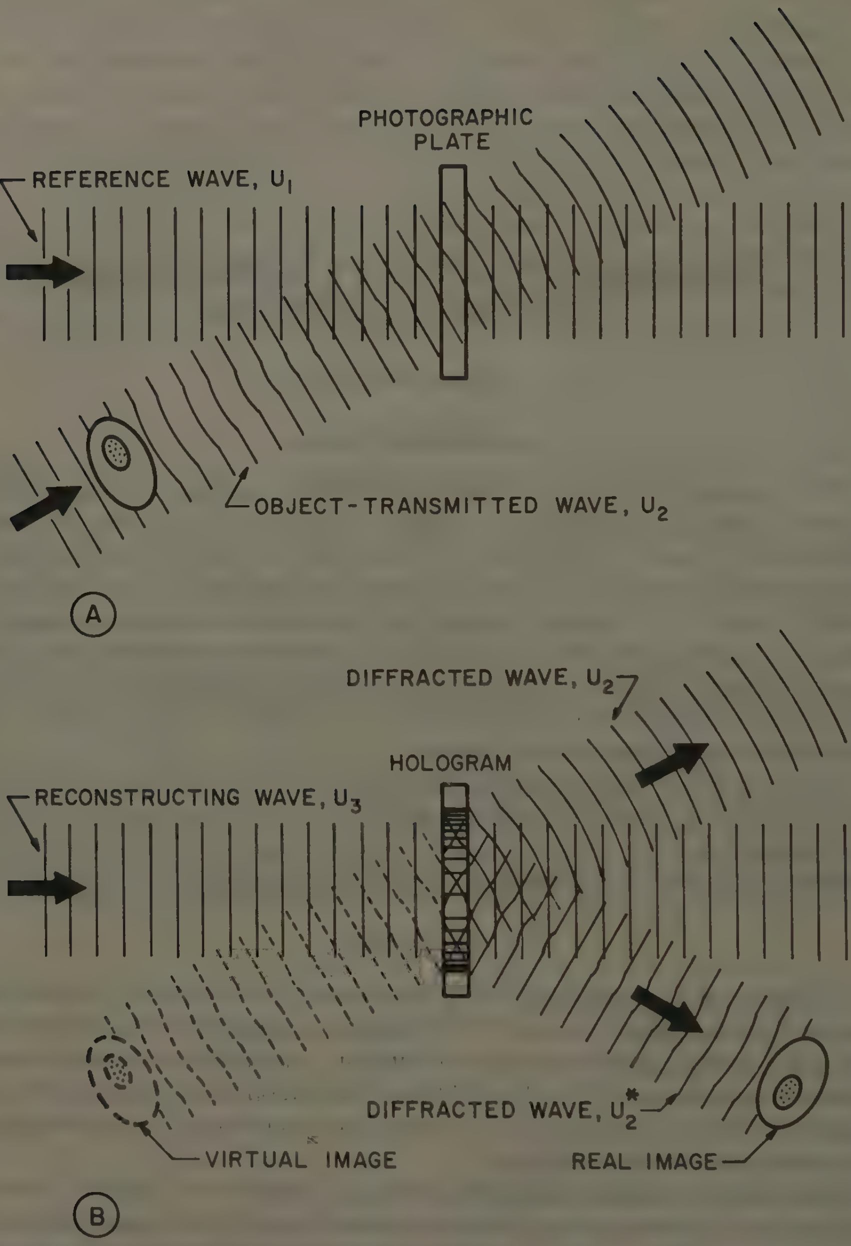

PHOTOGRAPHIC PLATE
REFERENCE WAVE, U_1
OBJECT-TRANSMITTED WAVE, U_2
A
DIFFRACTED WAVE, U_2
HOLOGRAM
RECONSTRUCTING WAVE, U_3
DIFFRACTED WAVE, U_2^*
VIRTUAL IMAGE
REAL IMAGE
B

Fig. 1. The upper figure (a) shows how the superposition of a reference wave and the object-transmitted wave forms the interference recorded on the photographic plate. The lower figure (b) shows how the recorded interference diffracts a reconstructing wave and forms the original object-transmitted wave, U_2, and its complex conjugate, U_2^*.

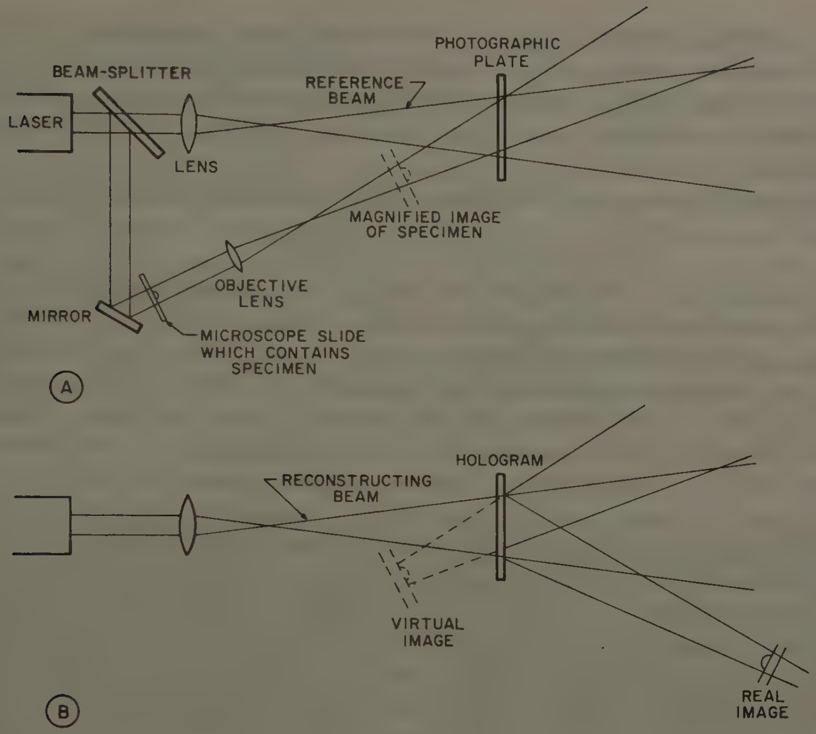

BEAM-SPLITTER
LASER
LENS
REFERENCE BEAM
PHOTOGRAPHIC PLATE
MAGNIFIED IMAGE OF SPECIMEN
OBJECTIVE LENS
MIRROR
MICROSCOPE SLIDE WHICH CONTAINS SPECIMEN
Ⓐ
RECONSTRUCTING BEAM
HOLOGRAM
VIRTUAL IMAGE
REAL IMAGE
Ⓑ

Fig. 2. A simple system for holographically recording microscopic specimens utilizes a high-quality microscope objective to form a magnified image of the specimen. The magnified image acts as the object for the hologram. During reconstruction, the recorded wave is formed producing a virtual image of the magnified specimen, and the complex conjugate of the recorded wave is formed producing a further magnified real image of the magnified specimen.

$$I = |U_1 + U_2|^2 = |U_1|^2 + |U_2|^2 + U_1^*U_2 + U_1U_2^*,$$

where the asterisk indicates complex conjugate. By keeping the bias of the recording high compared to the signal, the recording becomes effectively the intensity times some constant. Here the bias or d.c. terms are $|U_1|^2 + |U_2|^2$ and the signal or a.c. terms are $U_1^*U_2 + U_1U_2^*$. The bias is made large compared to the signal by making $|U_1|$ large compared to $|U_2|$.

Illumination of the processed recording with a third wave, U_3, results in spatial modulation of U_3 by the recording (Fig. 1-b.). The wave passed by the hologram is then, considering only the signal terms,

$$U_4 = U_3U_1^*U_2 + U_3U_1U_2^*.$$

In image formation, U_3 is made equivalent to U_1. Usually, both U_1 and U_3 are plane; thus $U_3U_1 = U_3U_1{}^* = C$, where C is a constant. Reconstruction of the object-scattered wave then occurs because $U_4 = CU_2 + CU_2^*$. In addition to the original object-scattered wave, CU_2, the complex conjugate of that wave,

CU_2^*, is formed. When U_1 and U_3 are not plane, but are identical, spherical waves, U_2 will be reconstructed with unity magnification. However, U_2^* will be formed with some change in curvature, that is, with magnification other than unity. By changing the curvature of U_3 from that of U_1, magnification of the reconstructed images can be varied.

The expression for magnification can readily be derived and shown to be (6)

$$M = \pm(1 - z_2 z_1^{-1} \pm z_2 z_3^{-1})^{-1},$$

where z_1, z_2, and z_3 are respectively the perpendicular distances from the reference, object, and reconstructing sources to the hologram. Positive M denotes a real image, and a negative M denotes a virtual image. If the object cannot be considered a single point, the magnification of different points in the object may differ. Thus, one or both images may be distorted in some apparatus configurations. The distortion can be predicted by using this expression for magnification.

Resolution of a hologram also depends on the geometry of the recording apparatus. As in other image-forming systems, resolution depends upon the aperture of the system. In holography, the aperture may be determined by the physical dimensions of the recording or by the ability of the medium to resolve the interference fringes. The spacing between fringes is $S = \lambda/(\sin \Theta_2 - \sin \Theta_1)$ where λ is the wavelength of the illumination and Θ_1 and Θ_2 are the angles of incidence of the two interfering waves. Portions of the hologram in which ε, the medium's resolution, is greater than S, the fringe spacing, will not record the fringes and will not contribute to the image during reconstruction. In the simple case of a small, axial object and an axial reference source, the hologram's resolution is (6)

$$\xi \simeq \lambda\epsilon/(\lambda + \epsilon \sin \Theta_{1-\max}).$$

Another limiting case is that where the object and reference source are close together compared to their distance from the recording plane. In this case, $\Theta_1 \approx \Theta_2$ over the recording plane so that the fringe spacing can be greater everywhere than the medium's resolution. Thus, the entire hologram contributes to the image and resolution is limited by the physical dimensions of the hologram. For a circular hologram,

$$\xi \simeq \lambda/\sin \Theta_H,$$

where Θ_H is the half-angle subtended by the hologram at the object.

Three-dimensional, phase-faithful images can have great value in biomedical research. Some obvious applications are studies of the interior of the eye, of capillary bed structure, of phagocyte migrations, or of cell division and growth (7). The phase-fidelity of holographic images permits the application of phase-contrast, dark-field, and interference techniques to extract additional information which is contained in the hologram, but not available in a bright-field reconstruction. Interference can be obtained in a variety of ways, for example, by superimposing holographic images, by superimposing a plane wave upon a holographic image, or by superimposing the object-scattered wave upon a holographic image (8). Superimposing waves during reconstruction permits the variation of superposition geometries in order to obtain a useful fringe spacing or orientation. Interference can also be obtained by multiply, usually doubly, exposing the hologram (Fig. 3) for example, by making two exposures of the specimen, allowing time for growth between exposures (6-7). Alternatively, one exposure can be made of the specimen and a second of a specified interfering wave (9) such as a plane wave, or one exposure can be made of the specimen in a relaxed state and a second of the specimen in a stressed state. Interference obtained by multiple exposures requires a simpler apparatus than interference obtained by superposition during reconstruction, but does not permit variation or adjustment of the fringe system and cannot be reversed to provide a bright-field image. In either method, a time variable is added to the spatial variables of conventional interferometry. This addition of a temporal factor to interferometry is a unique property of holography.

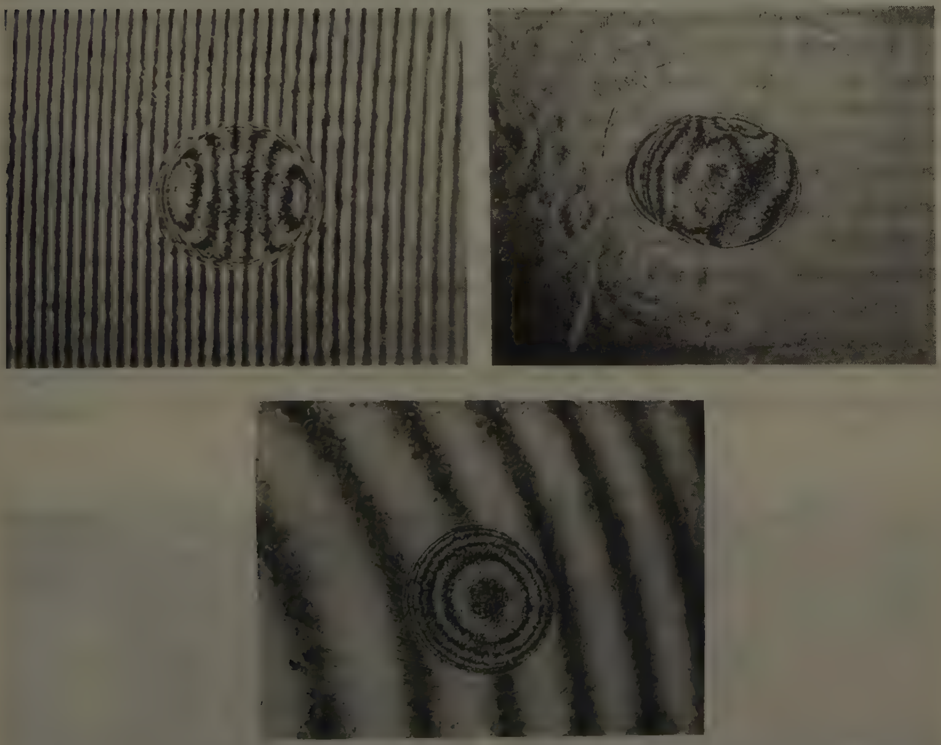


Fig. 3. Interference is illustrated in images of the eggs of the sea urchin, *Lytechinus pictus*. In each case, interference is obtained by doubly exposing the hologram during recording. At the left (a) is differential interference arising from a small angular displacement between exposures. At the center (b) is differential interference arising from changes occurring between exposures separated by about seven minutes. (The fertilized egg has begun to divide.) At the right (c) is infinite-fringe, duplication interference produced by removing the specimen between exposures. (The fringes correspond to contours showing the spherical shape of the egg.)

An interesting extension of holographic, interferometric techniques is the analysis of vibratory motion (10). If the specimen moves during exposure so that the pattern recorded by the hologram moves, the contrast of the recorded pattern will be reduced. If the contrast is sufficiently reduced, no image-forming diffraction can occur during reconstruction. A vibration specimen having a constant vibration amplitude will produce stable patterns only in its displacement maxima. Thus, the hologram will reconstruct images of the object in its maximal displacement configurations, and an interference pattern will be formed identical to that which would be formed by superimposing images of the specimen in these configurations.

Nonoscillatory motion also can be holographically detected (Fig. 4.). Again, the reduction of fringe contrast effected by a moving specimen causes a reduction of image intensity. When only a localized region of the specimen moves, only the corresponding region of the image will be darkened (6,7,11). This phenomenon can be the basis of detection and localization of very subtle motion or of motion of systems consisting of identical particles.

The fundamental equations of holography simply require that a stable interference pattern be maintained during recording. Interference is a general wave phenomen and can be obtained with a variety of types of radiation. Ultraviolet, infrared, microwave (12), or acoustical (13) radiation might be used to record the hologram and visible light used during reconstruction. Thus, a visible image is obtained using nonvisible illumination. Ultrasonic holography can be extremely valuable in visualizing soft tissue conditions in gynecology, obstetrics, ophthalmology, and tumor detection in general (6,18).

The phase and amplitude information necessary to form a hologram need not be derived by superimposing coherent reference and object-scattered waves. Alternatively, given a real or hypothetical, three-dimensional structure, the spatial intensity pattern of a corresponding hologram can be calculated (14,15). Naturally,

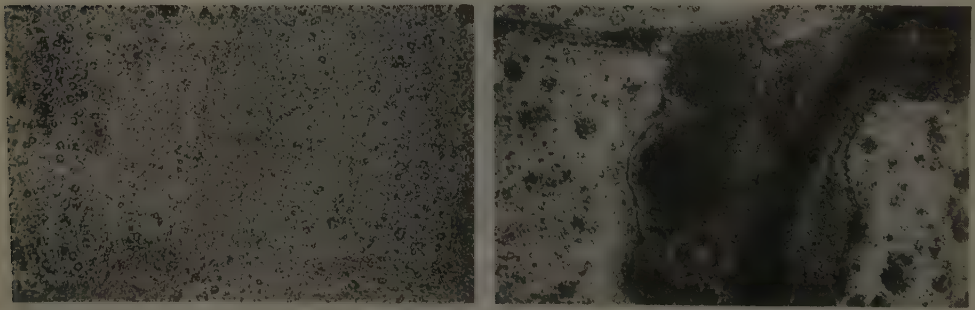


Fig. 4. Motion-induced contrast is demonstrated using a slime mold, *Physarum connatum.* The upper photograph (a) shows the mold, magnified about 200 times, as seen in coherent light. The lower photograph (b) shows the same region of the specimen with identical magnification as seen in a holographic image. The moving cytoplasm has darkened only localized regions of the image and has greatly increased the brightness range of the image.

modern computers are ideally suited to synthesizing such holograms. The computer can utilize a great deal of data corresponding to a large number of points in the object and convert this information into a two-dimensional function from which a hologram can be produced. Computer-generated holograms can be useful in visualizing medically significant three-dimensional structures. For example, ultrasonic techniques are currently used to study soft-tissue structures. Interfaces between tissues of different acoustical refractive index cause reflections of an acoustic pulse. The time of the pulse return is a measure of depth; the strength of the echo is a measure of refractive index differences. However, conventional oscilloscope displays lose a great deal of information because of their limited resolution and grey scale. The time and strength of the echoes might better be utilized in a holographic display. In this application, scans could be made in several, parallel planes to provide the computer with three-dimensional data. From this data, a hologram could be produced which would display ultrasonic echoes as an intensity distribution in three spacial coordinates. Diagnosis would be facilitated by direct, visual examination of internal, soft-tissue structures.

FILTERING

In addition to forming or synthesizing images, holograms can be used as filters to alter or analyze images (16,17). The fundamental equation of holography states that

$$U_4 = U_3U_1{}^*U_2 + U_3U_1U_2{}^*.$$

Again, for simplicity, only the signal terms are considered. The variables U_1, U_2 and U_3 in this expression can be considered the Fourier transforms of the variables, V_1, V_2 and V_3, so that

$$\begin{aligned} U_4 &= \mathfrak{F}[V_3]\cdot\mathfrak{F}[V_1{}^*]\cdot\mathfrak{F}[V_2] + \mathfrak{F}[V_3]\cdot\mathfrak{F}[V_1]\cdot\mathfrak{F}[V_2{}^*] \\ &= \mathfrak{F}[V_3\otimes V_1]\mathfrak{F}[V_2] + \mathfrak{F}[V_3\odot V_1]\mathfrak{F}[V_2{}^*]. \end{aligned}$$

In this expression, $\mathfrak{F}$ denotes Fourier transformation, $\otimes$ denotes correlation, and $\odot$ denotes convolution. In a real optical system, Fourier transformation is effected by a lens. If the complex amplitude distribution described by $V(x',y')$ is located at one focal pane of a lens, the complex amplitude described by $U(x,y)$ will be produced in the other focal plane. Thus, a hologram can be formed from V_1 and V_2 which contains the signal terms $U_1{}^*U_2 + U_1U_2{}^*$ (Fig. 5). This hologram, illuminated with U_3, by transforming V_3, will modulate U_3 to form U_4. In turn, U_4 can be transformed by a lens to produce V_4 where

$$V_4 = (V_3\otimes V_1)V_2 + (V_3\odot V_1)V_2{}^*.$$

Thus, we can obtain a virtual image consisting of V_2 times the correlation of V_3 and V_1, and a real image consisting of $V_2{}^*$ times the convolution of V_3 and V_1. Because of the intensity of the reconstructed V_2 depends upon the correlation or the convolution of V_3 and V_1, the reconstruction intensity can be used as a measure of correlation or convolution. A more useful application of this filtering property might be the detection of complex amplitude distributions which are

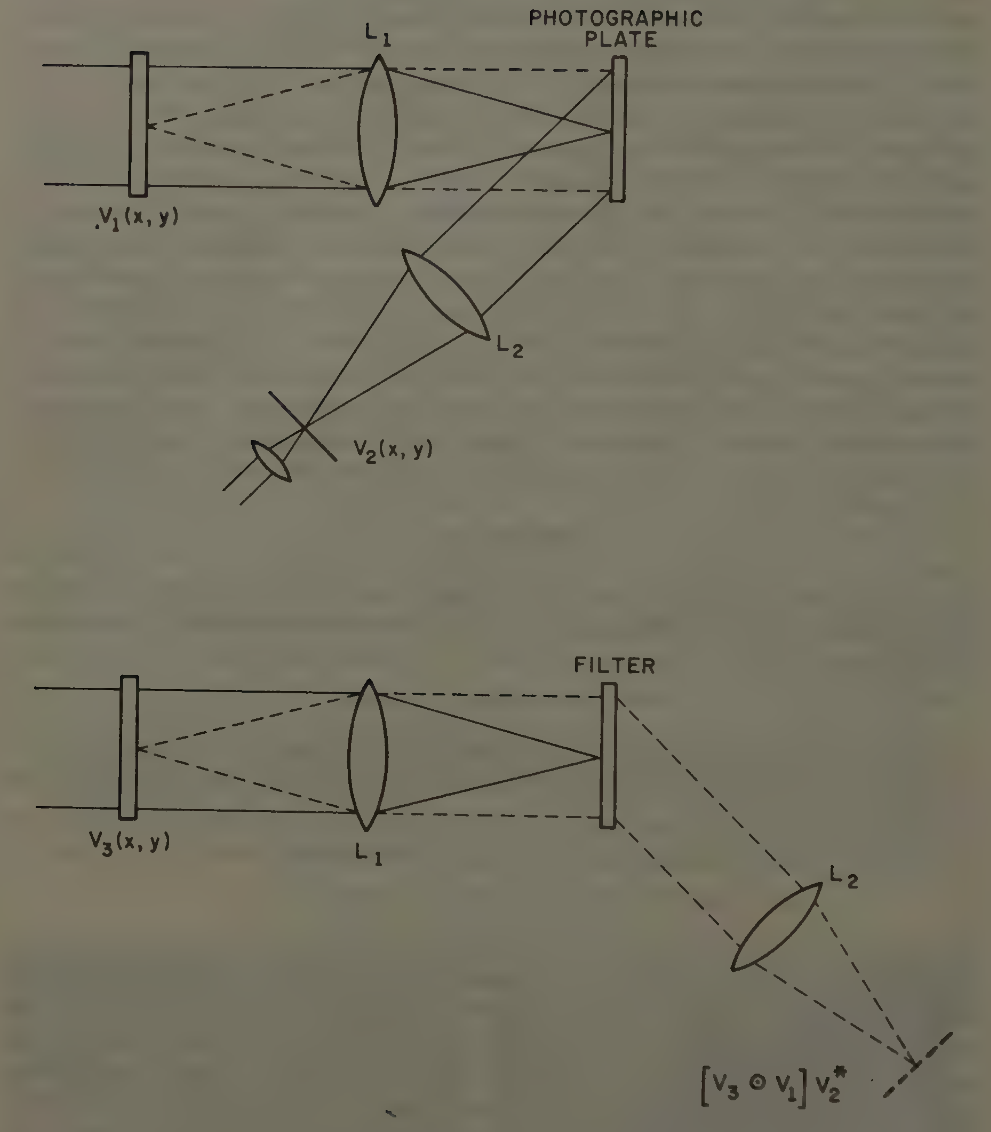


Fig. 5. Correlation and convolution filters are formed in Fourier space. The upper illustration (a) shows the recording of the filter from the Fourier transform of the reference pattern, $V_1(x,y)$. The lower illustration (b) shows the analysis of the test pattern, $V_3(x,y)$ by the filter.

related to the distribution of V_1 (17). In this case V_2 is most simply made a point source and V_1 is the reference specimen. The hologram made from the transform of V_1 and V_2 is then illuminated by the transforms of an array of patterns to be analyzed. A set of image points will be reconstructed in the focal plane of the output lens, each point corresponding spatially to a test pattern in the array. The intensity of each point will be a measure of the correlation or convolution of its corresponding test pattern with the reference pattern, V_1.

An outstanding potential medical application of holographic correlation/convolution analysis is the recognition of malignant cells in cytological diagnosis (6,18). Normally such diagnosis requires tedious scanning of a specimen on a microscope slide. Malignant cells must be identified against a dominant background of benign cells. A holographic filtering technique might be used to locate suspicious cells, thereby significantly reducing the tedium, and resulting risk of error, in cytological diagnosis (Fig. 6.).

Holographic filters of a different type can be useful in improving image quality or information content (19,20,21). The image can be made to more closely approximate an ideal or geometrical-optical image. Real optical systems form an image which is the convolution of the geometrical-optical image and a function which describes the departure from the geometrical-optical image. This function is called the spread function of the system. Using the V notation introduced above, the relationship is

$$V_3 = V_0 \odot V_2,$$

where V_3 is the actual image, V_0 is the ideal or geometrical image, and V_2 is the spread function. The spread function is the image which the optical system will produce from a point object. In the Fourier domain the above expression becomes

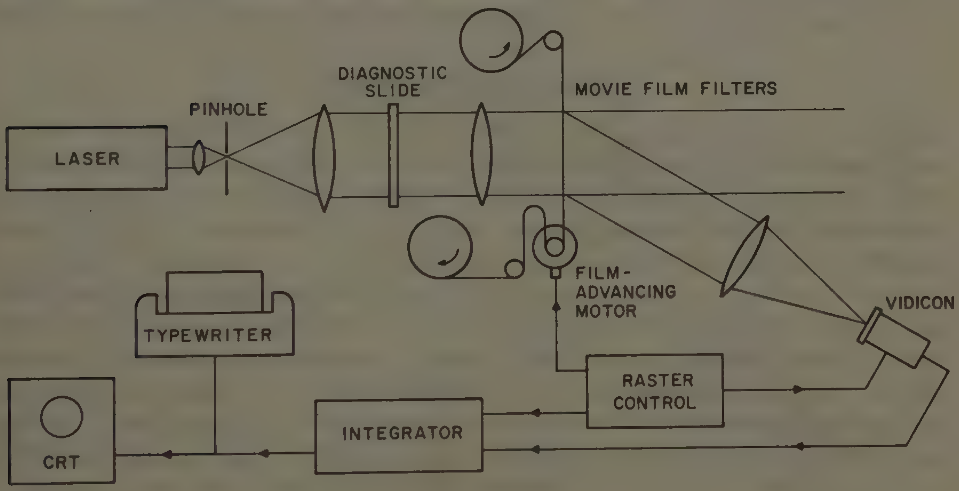


Fig. 6. Cytological diagnosis might be aided by a holographic, convolution-filtering method. Filters produced from benign and malignant cells are recorded on individual frames of a movie film. The transform of the diffraction pattern produced by the diagnostic slide is filtered frame by frame. In the output plane, the filtered wave is scanned by a vidicon and the output integrated in two dimensions for each frame. The final output of the system is the coordinates of cells which resemble malignant cells.

$$U_3 = U_0 U_2,$$

where U_2, the Fourier transform of V_2, is the modulation transfer function or MTF. If the inverse of U_2 could be obtained, modulating U_3 with U_2^{-1} would give $U_3 U_2^{-1} = U_0 U_2 U_2^{-1} = U_0$. In fact $U_2^{-1} = U_2^* |U_2|^{-2}$ can be obtained by a combination of holographic (or complex amplitude recording) and photographic (or intensity recording) techniques. It is obtained from a recorded image of a point source produced by the optical system in question. The record (e.g., photographic transparency) is illuminated by a plane, coherent wave (Fig. 7.). The transmitted wave, V_2, is transformed by a lens, providing U_2. If U_2 is simply recorded photographically, the photograph contains $|U_2|^{-2}$. If a reference wave, U_1, is superimposed upon U_2, a hologram containing $U_1^* U_2 + U_1 U_2^*$ is formed. By making U_1 plane, this hologram essentially becomes $C(U_2 + U_2^*)$. The photograph containing $|U_2|^{-2}$ and the hologram containing CU_2^* are combined to provide the desired filter $U_2^* |U_2|^{-2}$. Illuminating this filter with U_3, the transform of V_3, causes U_3 to be modulated by U_2^{-1}, providing U_0. A lens transforms U_0 producing the geometrical-optical image, V_0.

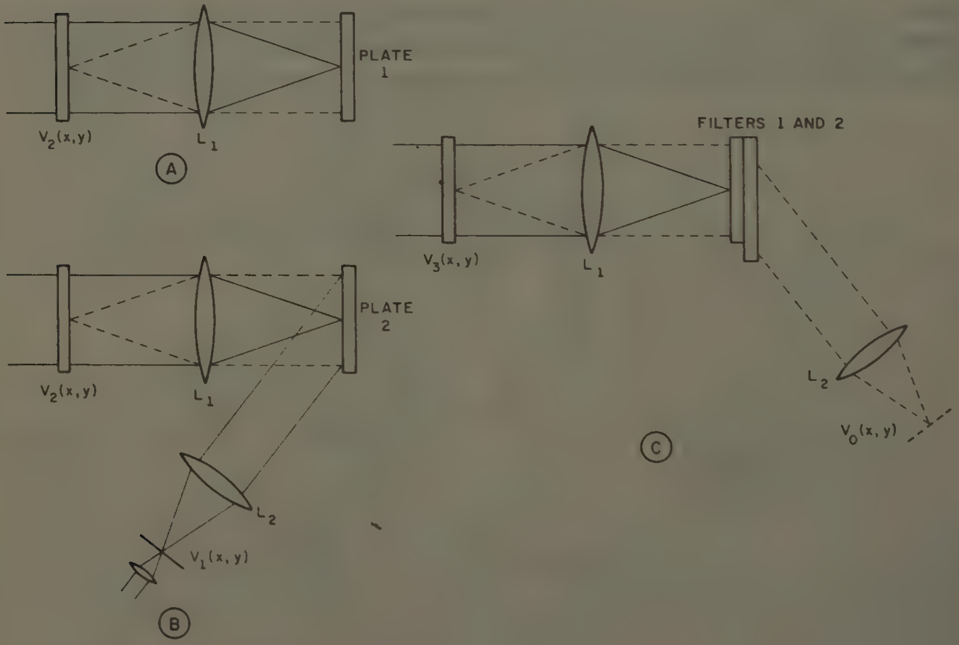


Fig. 7. Resolution-enhancing filters are made by (a) photographically and (b) holographically recording the MTF. The MTF is obtained by transforming the image $V_2(x,y)$, of a point source. Subsequently (c), the photograph to be sharpened, $V_3(x,y)$, is filtered by the sandwiched photograph and hologram. The wave passed by the filter is transformed to provide the sharpened image, $V_0(x,y)$.

A useful application of this "deconvolution" filter is the enhancement of the resolution and contrast of conventional radiograms (6,18). The spread function of a radiographic system is obtained by making a radiogram of a pinhole aperture. This radiogram, consisting of V_2, provides U_2 by the procedure described above. By making filters for a number of different aperture-to-radiogram distances, a number of different planes can be selectively sharpened in the diagnostic radiogram. Thus, three-dimensional information is obtained from a single, conventional, two-dimensional photograph. Moreover, since the dependence of resolution on a small source of X-rays is relaxed, larger sources can be used thereby allowing higher X-ray intensities and reducing the danger of blurring due to patient motion.

This deconvolution technique can be applied to any system for which a spread function, or its transform, the modulation transfer function, can be defined (22-23). Applied to isotope scanning systems, it can provide three-dimensional information from a single-plane scan. Large aperture gamma-ray cameras might be developed which can provide images of activity distribution. In these applications, the MTF need not necessarily be photographically recorded. In some situations, calculation (24) of the inverse of the MTF, or U_2^{-1}, directly from the spread function (22) may be a simpler method than optical filter formation. Computer calculation of U_2^{-1} would be especially appropriate when data is acquired digitally and stored and treated in a computer, as, for example, in conventional isotope activity scans. Computer deconvolution of the scan data can significantly enhance the resolution of the scanning system and simultaneously provide three-dimensional data from a two-dimensional scan.

GENERALIZATION

We have been discussing holography as an image-forming technique and as an image-analyzing technique. In these applications holography can be defined as a complex, spatial, optical filtering technique. However, we can even more generally define holography as an optical, signal-processing method. Notice that this is not written as "a method for processing optical signals." The ultimate derivation of the signal need not be optical. Electrical signals have been analyzed using holographic methods (25). While originally the electrical application was pulse-radar echoes, the electrical signal might be an electro-cardiogram or electroencephalogram. The signal could have a nonelectrical derivation, for example, a fetal heartbeat, respiratory sounds, or eye movements. Holographic correlation/convolution methods might be used to recognize and distinguish a relevant signal from irrelevant, background signals, or to analyze the relevant signal for diagnostic purposes.

We can extend our generalization one step further and define holography as an information processing concept. Holograms might be used as optical memory or computing elements in computers. The capability of holograms in performing two-dimensional correlation and convolution has been discussed. However, holographic theory may have even more abstract, fundamental, conceptual applications. For example, holographic theory has been utilized in the study of associative memory (26). Modern theories of memory maintain that chemical substances are responsible for memory. Presumably experiences effect the metabolism of one or many chemical species, and presumably this effect is spatial. Holographic concepts are implicated in the determination of the spatial distribution of these substances. New experiences are related to prior experiences in a manner analogous to the relationship of the object-scattered wave to the reference wave. In associative memory, the "coherent" superposition of the new experience with previous experiences produces a cortical "interference pattern" in the brain. The relationship among experiences is determined by this pattern, and an experience may be "reconstructed" by illuminating the "recording" with a "reference" experience.

CONCLUSION

Holography is far more than a technique for producing three-dimensional images. Rather than being a specific technique, holography is a versatile concept encompassing many techniques. While holography originally might have been defined as an interferometric wavefront recording technique, it soon developed into a more general, complex, spatial filtering technique, and then into an even more general signal processing method. Recently, the potential of holographic theory in the description of associative phenomena has been demonstrated.

Medical research can profitably utilize holography in a range of applications. Imaging techniques provide images retaining full depth information, and which can be manipulated to provide dark-field, phase-contrast, or interference images. Image formation depends on the recording of stable fringes so that image intensity depends upon object stability. Thus, localized, moving systems containing indistinguishable, slowly moving or very small particles can be detected and analyzed. Holograms, providing visible images, can be recorded using nonvisible illumination. These contain all the attributes of holographic images while being formed in microwave, infrared, ultraviolet, or acoustical illumination.

Nonimaging applications of holography can also be valuable in medical research. Correlation filters can analyze the correlation of various specimens. One application of this technique is the reduction in time required to examine diagnostic cytological specimens. Resolution retrieval filters can provide three-dimensional information from two-dimensional images. Furthermore, the resolution of conventional techniques, X-ray diagnosis for instance, can be significantly improved. Signals which are a function of time also can be analyzed by these methods. Electrocardiograms and electroencephalograms are obvious medical examples. Holographic theory may even prove useful in a conceptual sense in describing spatial phenomena such as associative memory.

REFERENCES

1. D. Gabor, *Nature* **161,** 777 (1948).
2. D. Gabor, *Proc. Roy. Soc.* (London) **A 197,** 454 (1949).
3. E. Leith, J. Upatnieks, *J. Opt. Soc. Am.* **52,** 1123 (1962).
4. E. Leith, J. Upatnieks, *J. Opt. Soc. Am.* **53,** 1377 (1963)
5. E. Leith, J. Upatnieks, *J. Opt. Soc. Am.* **54,** 1295 (1964).
6. E. Feleppa, *Physics Today* **22,** No. 7, 25 (July 1969).
7. E. Feleppa, Ph.D. Dissertation, Columbia University (1968).
8. R. Brooks, L. Heflinger, R. Wuerker, *Appl, Phys, Lett,* **7,** 248 (1965).
9. L. Heflinger, R. Wuerker, R. Brooks, *J. Appl. Phys.* **37,** 642 (1966).
10. R. Powell, K. Stetson, *J. Opt. Soc. Am.* **55,** 1593 (1965).
11. E. Feleppa, *J. Cell Biol.* **40,** 838 (1969).
12. E. Rope, G. Tricoles, *J. Opt. Soc. Am.* **57,** 96 (1067).
13. A. Metherell, E. El-Sum. J. Reher, L. Larmore, *Appl. Phys. Lett.* **10,** 277 (1967).
14. A. W. Lohman and D. P. Paris, *Appl. Opt.* **6,** 1739 (1967).
15. L. Lesem, *ACM Commun.* **11,** 661 (1968).
16. D. Gabor, G. Stroke, R. Restrick, A. Funkhouser, D. Brumm, *Phys. Lett.* **18,** 116 (1965).
17. G. Stroke R. Restrick, A. Funkhouser, D. Brumm, *Phys. Lett.* **18,** 274 (1965).
18. E. Feleppa, in "Advances in Biomedical Engineering and Medical Physics," Vol. 4, ed. S. Levine (to be published) (1970).
19. G. Stroke, R. Zech, *Phys. Lett.* **25,** 89 (1967),
20. A. Lohman, H. Werlich, *Phys. Lett.* **25,** 570 (1967).
21. G. Stroke, *Phys. Lett.* **27,** 405 (1968).
22. T. Cradduck, These Proceedings (Symposium D.).
23. K. Doi, *Am. J. Roentgenol. Rad. Therapy and Nuclear Med.* **94,** 712 (1965).
24. A. Lohman, D. Paris, *Appl. Opt.* **7,** 651 (1968).
25. M. Arm, M. King, *Appl. Opt.* **8,** 1413 (1969).
26. K. Pribam, *Sci. Am.* **220,** 73 (1969).

CARDIAC ASSISTANCE BY THE INTRA-AORTIC BALLOON

A. R. Kantrowitz, P. N. Madras, J. D. Laird,
R. T. Jones, W. G. Austen and M. J. Buckley

The heart is one of the few organs which has a basically simple, mechanical function, namely to pump blood. In spite of this, the treatment for heart failure has long consisted of biochemical and pharmacological intervention with the heart and other organ systems. Recently, however, mechanical forms of treatment for heart failure have been gaining acceptance. In this presentation we will review some of the pertinent history of one form of mechanical cardiac assistance called counterpulsation, followed by a discussion of some of the work which our laboratory has performed to develop a cardiac assist system.

Some of the initial concepts leading to counterpulsation were published in 1953, (1) when Kantrowitz and Kantrowitz proposed that one could improve cardiac function by increasing the pressure at the orifices to the coronary arteries during diastole, since it is only at this time that significant flow through these vessels occurs. In the early 1960's, Clauss et al. (2) extended this idea to simultaneously lowering aortic pressure during systole, thereby diminishing the work which the heart must perform to eject its volume.

The means to accomplish this reduction in aortic pressure during systole and elevation in diastole has been highly varied. Clauss (2) originally did this by withdrawing blood from the aorta starting immediately before systole, allowing sufficient time for the reduced pressure to propagate as a wave up the aorta to the heart. Then he reinjected the blood starting just before diastole in order to raise the aortic pressure. This system to inject and withdraw the blood was used initially on patients in the early 1960's, but fell out of favor until more recently (4). It has many disadvantages. In order to move sufficient blood volume, high flow velocities and high shears must be used. These shears harm the formed elements in the blood, thus limiting the duration of use of this device. Furthermore, these shears could also damage the diseased blood vessel wall through which blood is pumped.

Finally, the blood vessels to the legs, which are used for insertion of the cannulas through which the pump is attached, are completely occluded thus stopping blood flow to the legs, a condition which can be tolerated for only a few hours.

Birtwell and Soroff (3) also developed an external means of producing these pressure changes by placing the patient's legs in a pressure jacket which could squeeze the legs at the appropriate times. By squeezing the legs, blood is forced out of them into the aorta, thus increasing the aortic pressure. Similarly, reducing pressure on the legs will drop the aortic pressure. This device offers the advantage of requiring no surgical intervention in the ill patient.

One effective means of producing the pressure changes was the auxiliary ventricle (4). This device was permanently implanted by a left thoracotomy in a three hour surgical procedure. The patient was not necessarily pumped continuously. While dramatic clinical improvement was noted when the device was used, it was very prone to form lethal thrombi so the device was clinically unacceptable.

Early in the decade, Moulopoulos, Kolff and Topaz (5) suggested that the required pressure changes could be produced by means of a gas driven balloon inserted high in the aorta (Fig. 1). This device had the advantage of simple, atraumatic insertion, tolerable shear levels in the blood, and adequate circulation to the limbs. After initially trying this balloon on a few patients, they reported some beneficial effects but no follow-up studies were performed. Since the device appeared simple with no obvious drawbacks, we decided to develop an intra-aortic balloon for clinical use.

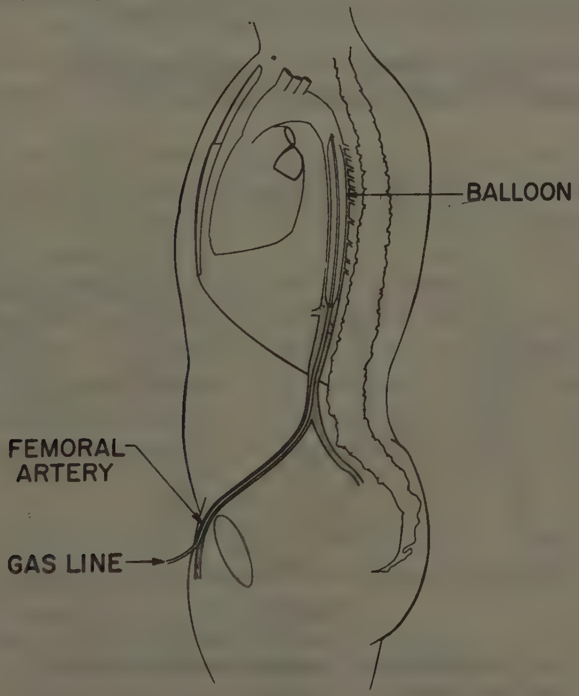


Fig. 1. Placement of the Intra-Aortic Balloon in the Descending Thoracic Aorta.

Principle of Counterpulsation

The principle of increasing pressure in the aorta during diastole and decreasing pressure during systole is known as counterpulsation. To accomplish this, the R wave is detected by electrodes used for patient monitoring. The R wave is used to start two timing circuits, one which times the initiation of balloon inflation, and the other which sets the duration of inflation. These times are set so that the pressure rise caused by inflation will reach the aortic valve just after its natural closure, so that high pressure is felt at the coronary arteries but not while the heart is ejecting blood. The pressure drop caused by deflation appears at the aortic valve when the left ventricle is starting to contract, so that less pressure is needed by the heart to expel its volume. The effect of these pressure changes on the heart may be defined in the more precise manner if two models for the heart are examined. In Fig. 2 we assume the heart to be flow limited, the case of fixed cardiac output. The assumptions are also made that the arterial tree is a windkessel, that the balloon volume is equal to the heart stroke volume and that inflation and deflation of the balloon occur instantaneously. The windkessel approximation is useful for rough calculations. In a later formula, variation in stroke volume will be accounted for and the time required for inflation and deflation will be corrected in considerations presented below.

In Fig. 2 the reduction in left ventricular pressure is equal to the difference in aortic systolic and diastolic pressure. Furthermore, the pressure produced by balloon inflation in diastole is also equal to this pressure difference.

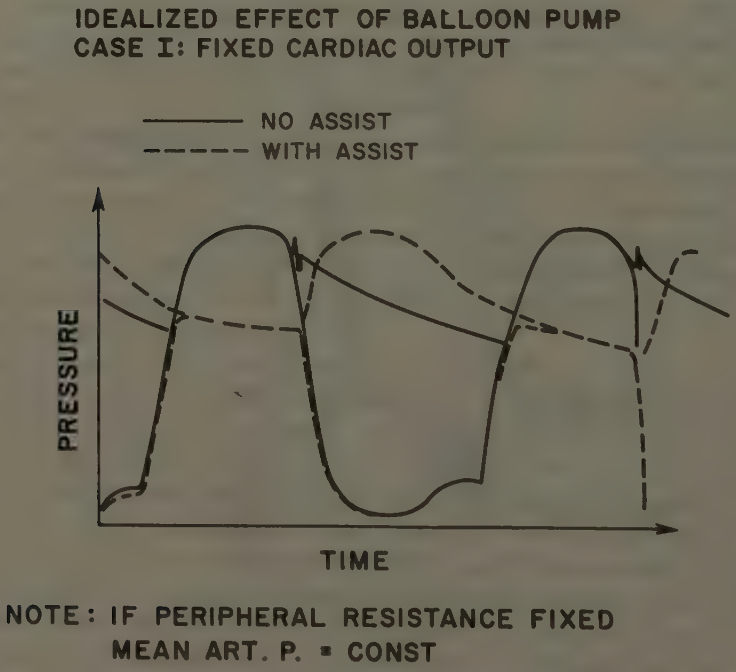
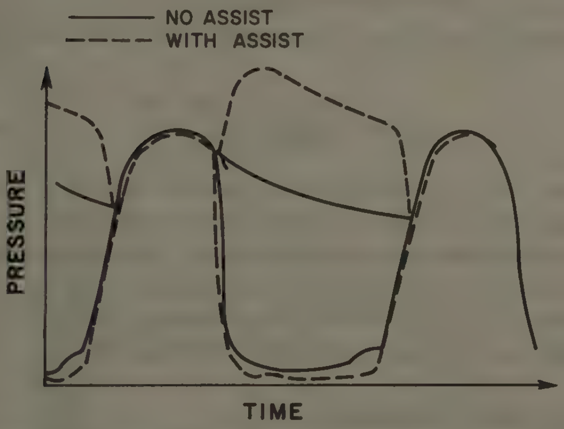


Fig. 2. Effect of Counterpulsation Assuming a Fixed Cardiac Output.

IDEALIZED EFFECT OF BALLOON PUMP
CASE II: FIXED PRESSURE VENTRICLE



NOTE: IF PERIPHERAL RESISTANCE IS CONST.
C.O. INCREASES, MEAN ART P. INCREASES

Fig. 3. Effect of Counterpulsation Assuming a Fixed Left Ventricular Pressure.

In Fig. 3 we make the assumption that the heart is limited by the pressure which it can produce. Thus, no reduction in left ventricular pressure occurs when the balloon is used, but a greater effect in diastolic augmentation pressure occurs. Thus, an overall rise in mean arterial pressure takes place due to an increased cardiac output.

In reality, the failing heart is probably closer to the fixed pressure approximation, and hence cardiac output increases when the balloon is pumped. This finding is compatible with work done in 1958 by Sarnoff et al. (6) in which they found that the energy consumed by the heart muscle was proportional to the pressure which the muscle produced and not the volume pumped by the heart (and not the mechanical work performed by the left ventricle or the heart as a whole).

When the balloon is turned on, the heart functions closer to the fixed pressure model described above. However, some reduction of left ventricular pressure is also found. A formula describing the volume and pressure changes caused by the balloon, or any arterio-arterial blood pump, has been derived, by Jones et al. (7).

$$\frac{\Delta Q}{Qh} - \frac{\Delta P}{P_S} = \frac{Qb}{Qh} \times \frac{P_S - P_D}{P_S} \qquad (1)$$

where ΔQ = change in cardiac output per stroke—due to assist
ΔP = change in LVP—due to assist
Qh = left ventricular stroke volume
Ph = left ventricular pressure = P_S
Qb = balloon stroke volume
P_S = systolic pressure—without assist
P_D = aortic diastolic pressure—without assist

Since all these parameters are measurable in any experiment, the equation may be treated as a criterion for effective balloon pumping, or counterpulsation of any sort, using the assumptions mentioned above.

Comparing the L.H.S. of Eq. (1) with the R.H.S., we have the efficiency (η) for the pump

$$\eta = \frac{P_S \cdot \Delta Q - Qh \cdot \Delta P}{Qb(P_S - P_D)} \tag{2}$$

Using Eq. (2) we find (8) that balloon η never exceeds 70-80%. The "inefficiency" is accounted for by the assumption made earlier that no time is required for inflation or deflation. In fact, both inflation and deflation require from 50 to 70 msec. If the total period one has for balloon inflation is about 350 msec, which is the case for a heart rate of 100 beats/minute, then the inefficiency is accounted for by the time required to move gas in and out of the balloon.

Inflation Characteristics of Balloons

The preceding discussion is applicable, with minor modifications to any counterpulsation pump placed "in series" with the heart. The balloon presents a specialized set of problems which will be discussed in this section.

If a balloon has a larger inflated diameter than the diameter of the aorta, the resulting efficiency, measured by Eq. (2), will be about 10%. When the balloon is significantly smaller than the aorta, its efficiency will be at the predicted 70-80%, but its effect will be small because volume Qb is unnecessarily limited. The problem of scaling the balloon is even more complex because aortic diameter is very pressure dependent and it is at the lowest pressures that most pumping is necessary, since low pressures are the result of shock, the condition which the balloon is meant to treat. The significance of aortic elasticity in the patient is illustrated by the fact that aortic volume doubles between a shock mean of 30-40 mm Hg and a normal mean pressure of 90 to 100 mm Hg.

Ideally, one might design a balloon to occlude the aorta and hence pump all the available volume. Such a balloon could follow changes in aortic size with pressure and would be suitable for use in all patients. However, as mentioned earlier, an occlusive balloon only offers 10% efficiency. If one measures the pressure beside this balloon (Fig. 4), one finds that it is very high, approaching the levels which are used to drive helium into the balloon. Pressures measured at the ends of the balloon are very small, in marked contrast to the lateral pressures. Thus, the balloon must be inflating at the ends first as shown in Fig. 4.

This tendency for the balloon to inflate in this manner is easily understood. The material from which the balloon is made has no structural rigidity when the balloon is not completely inflated. As the balloon inflates, a column of blood having a density close to unity, i.e., one thousand times greater than the gas density, must be moved rapidly from the vicinity of the balloon. Thus, a pressure gradient must exist in this blood column, with the highest pressures occurring at the column center, and the lowest at the ends. Because for the unstretched balloon the internal and external pressures are equal, the gas in the balloon will flow to the

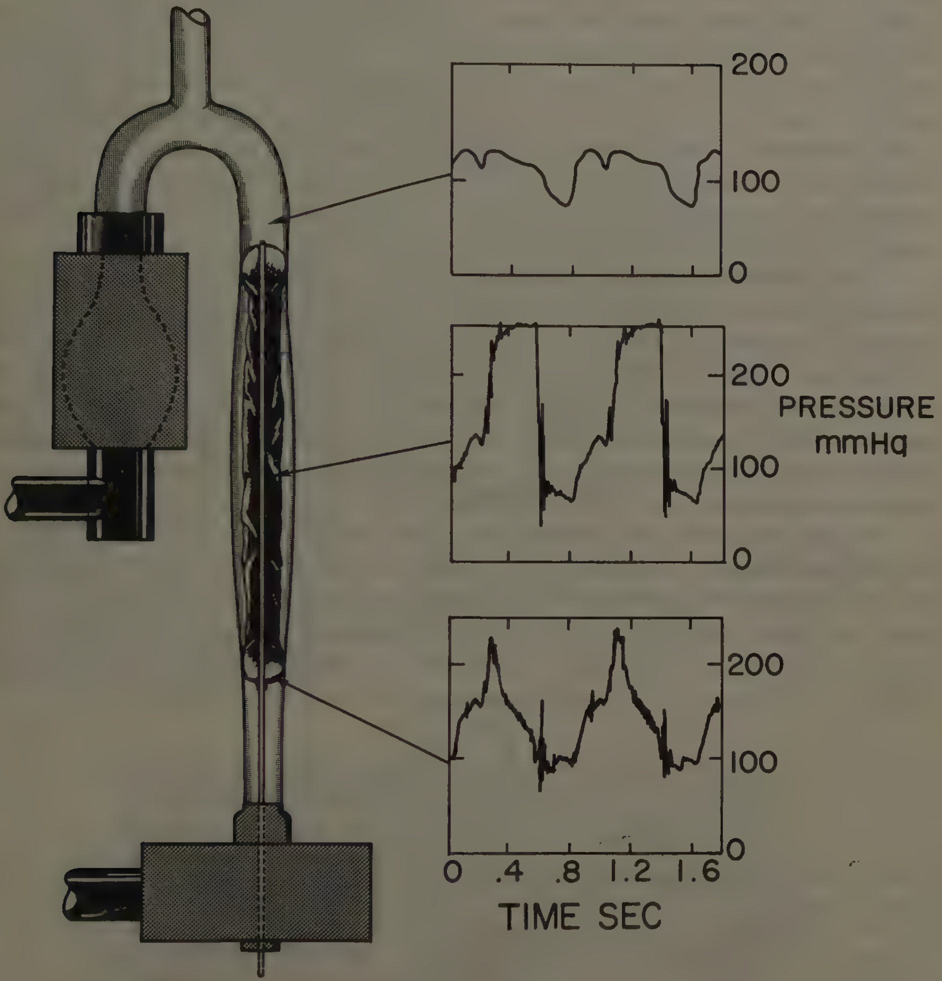


Fig. 4. Pressure Tracings Measured—Proximal, Lateral and Distal to the Intra-Aortic Balloon Placed in a Model Circulation System. Note the inflated ends of the balloon which prevent effective pumping. This is an illustration of "Bubble Blowing".

lowest pressure region at the ends of the balloon. Thus, the balloon ends must inflate first, and if they are capable of occluding the aorta, they seal the blood opposite the center of the balloon and prevent adequate pumping. This phenomena has been labelled "bubble blowing" (8).

The consequences of "bubble blowing" are of special importance to balloon design. Consider that the aorta is an elastic organ. Its inner diameter at 40 mm Hg = 14 mm, at 100 mm, and at 150 mm Hg = 20-24 mm (9).

These are mean figures for old aortae, typical of those in which a balloon might be used. Normal pressure is close to 100 mm Hg giving a pumpable volume of 65 cc. The patient with the balloon would probably have pressures near the normal range for most of the pumping duration. However, one cannot build a balloon for this mean pressure range since the patient may be hypotensive when the balloon is inserted, or at any point in the course of his treatment, and it is then in the hypotensive range that pumping would be sacrificed due to bubble blowing. It is when the patient's pressure is low that effective pumping is most urgently needed. Thus, the balloon must be made to pump in aortae at low pressure, even though this entails a sacrifice in efficacy at normal blood pressures.

To aid in overcoming the problem of bubble blowing, the balloon was divided into three segments of equal length. The gas is admitted most rapidly to the center segment, by making the orifices to this segment larger than the orifices to the two end segments (Fig. 5). Segmentation, however, is only a partial solution to the bubble blowing problem. Maximum pumping efficiency is retained with a larger segmented than unsegmented balloon. Nevertheless, the tendency to blow bubbles is still present so the balloon is also tapered slightly, and selected so as not to occlude the aorta into which it is to be placed.

Having determined for humans (via a search of scanty literature), and for pigs (by experiment) exactly what size balloon is necessary to pump effectively at various pressures, we performed experiments designed to test out balloons, both against the theory in Eq. (2) and against a displacement pump (an auxiliary ventricle) attached to the aorta at the same point as the proximal tip of the balloon (Fig. 6) (8). The results of this experiment are shown in Figs. 7 and 8.

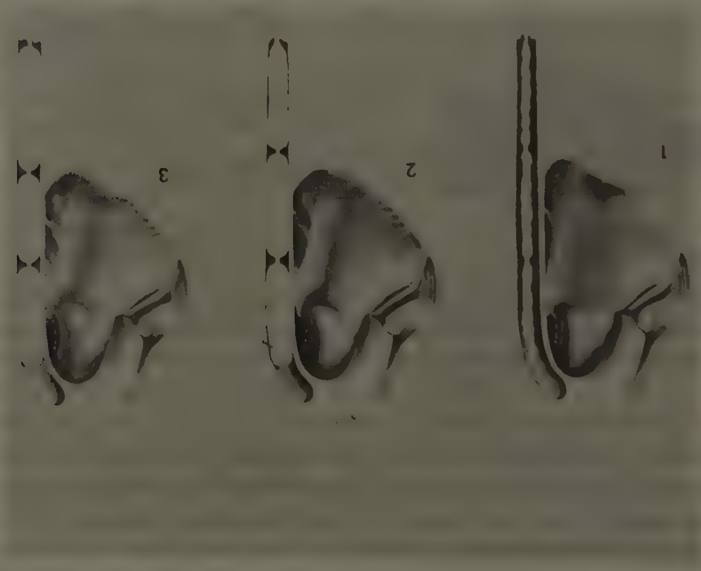


Fig. 5. Sequence showing inflation of the three segment balloon. The center segment inflates more rapidly than the two end segments, thus helping to circumvent the problem of bubble blowing.

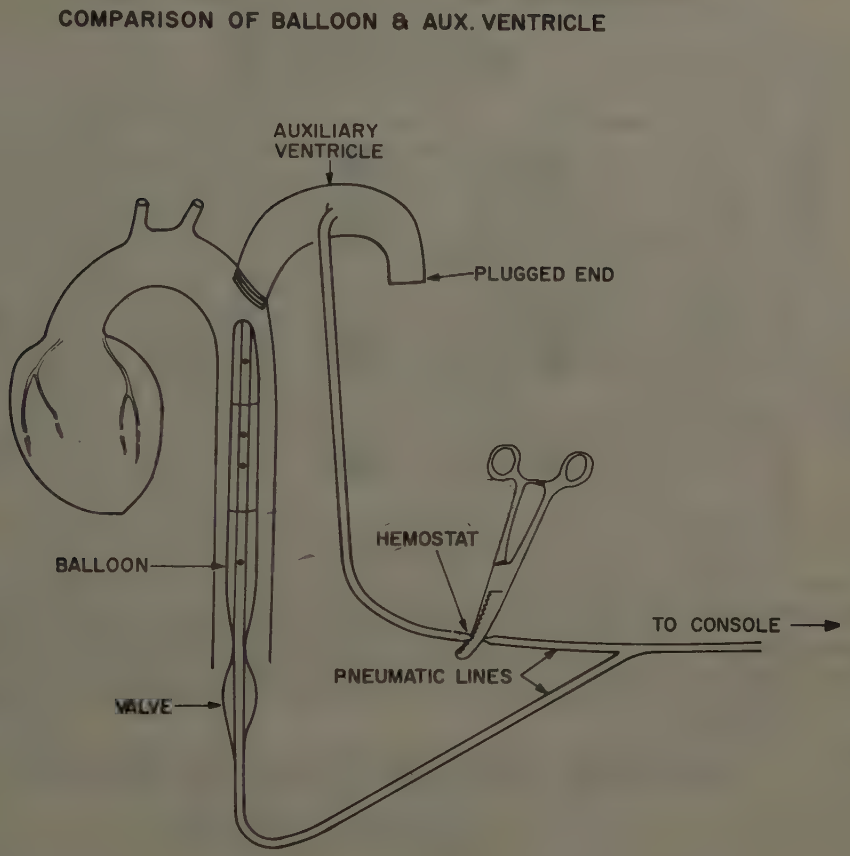


Fig. 6. Experiment in which the balloon and a displacement pump the "Auxiliary Ventricle" are placed in the same aorta. Pumping either device is accomplished by altering the position of the hemostat to block one of the pneumatic lines. The "valve" is a separately inflatable segment which occludes the aorta, thus elevating the mean aortic pressure proximal to the valve. This rise in pressure causes an increase in aortic diameter, thus allowing the experimenter to make a comparison of balloon and auxiliary ventricle performance at varying aortic diameters.

Safety Evaluation

At this point, we felt that the pumping efficacy of the balloon was well understood and quantifiable. Therefore, a clinical system was designed and built.

In order to make such a system usable in the hospital, many engineering requirements in the control unit were met.

Firstly, the electrical isolation of the patient from the console is an obvious demand on the engineering design. This problem is complicated by the fact that most patients in an intensive care unit will be connected to several other electronic instruments and the potential for ground loop hazards (11) is very real. Our

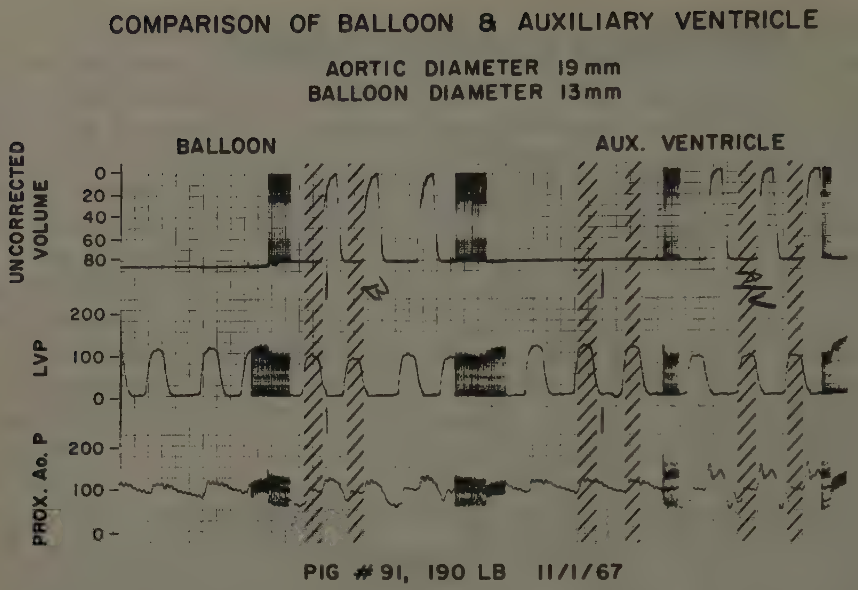


Fig. 7. Comparison of balloon and auxiliary ventricle in an aorta whose diameter is larger than the balloon diameter. The times when the balloon is pumping is indicated by a signal in the upper "uncorrected volume" channel. Note that both devices cause similar changes in left ventricular pressure and in diastolic augmentation pressure.

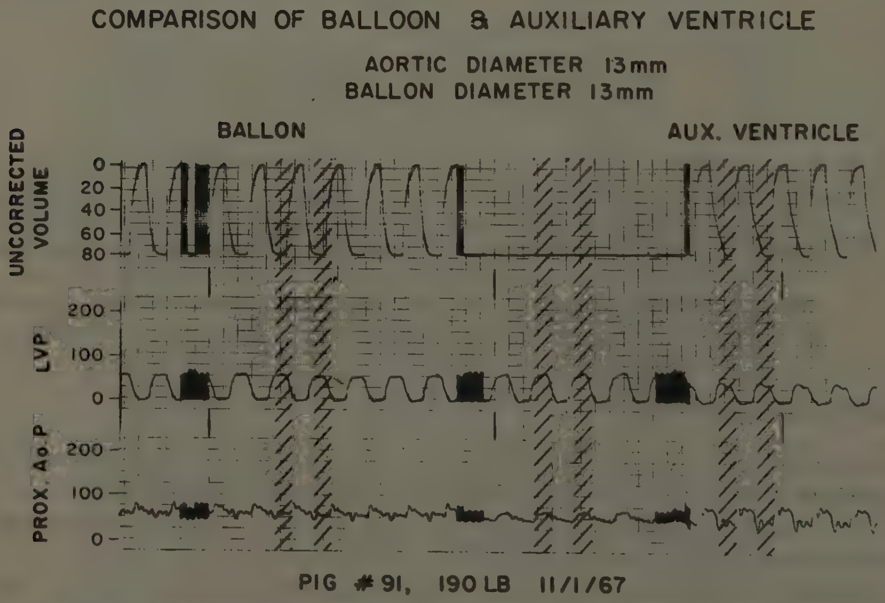


Fig. 8. In this case the aortic diameter is the same as the balloon diameter. While the effectiveness of the auxiliary ventricle is as great as in Fig. 7, the balloon exhibits almost no pumping capability.

approach has been to make no presumptions regarding the adequacy of grounding, line. The circuitry has been designed so as to make the chance of trauma to the patient very improbable. In addition, it is necessary that the electronics withstand interferences, etc. from such ancillary equipment as electro-cautery, defibrillators, etc. Secondly, demands specific to the balloon are also placed on such a system. Should the intra-aortic balloon burst, its entire volume would be injected into the arterial blood and would probably be fatal. This mode of failure, i.e., by massive rupture has never occurred with the material from which our balloons are made. Under much greater stress than these balloons encounter in human use, they fail in no less than fifteen weeks or fifteen times as long as they would be used in humans. The mode of failure has been the development of microscopic leaks which slowly grow. These leaks are small enough to be harmless at first, but they are detectable at a stage when they are less than 1/10 of the patient's ability to clear helium from his blood stream via his lungs. Thus, should such a leak develop in a clinical trial, the balloon could be replaced before damage is done to the patient.

An important hazard with any intravascular device is the formation of clots and emboli. We have developed a polyurethane copolymer which is very non-thrombogenic and which has adequate tensile strength, great flexibility, ease of handling and relative impermeability to gases. This material, Avcothane, was synthesized into various implantable forms, and these remained in the arterial system of pigs for up to eight months with no harmful effect. It has proven very satisfactory for balloon applications.

The problem of timing the balloon in relation to the cardiac cycle consists of two parts. The first is simply detection of the R wave, but the second part is much more subtle. The balloon is located about 15 cm from the aortic valve. The elastic characteristics of the aorta are such that the pressure or flow impulses propagate at a finite wave speed of about 500 cm/sec in the thoracic region. Hence if the full diastolic interval is to be effectively utilized for balloon pumping it is necessary that the propagation delay be accounted for in the setting of inflation and deflation times. For the typical values cited above the inflation should be started .030 sec prior to the time of aortic valve closure, if the pressure rise at the root of the aorta is to begin just after closure of the valve.

Similarly, the deflation wave must reach the root prior to the opening of the valve.

These delays must be accounted for when examining the pressure wave form as measured by the pressure transducer on the tip of the balloon catheter, since it is by that wave form or a radial artery pressure that the timing is "fine-tuned."

Using this system, we pumped seven farm pigs continuously for one week. These pigs were perfectly awake, unsedated, and kept in large cages with normal diet and activity. The pigs showed no signs of awareness that they were being pumped. They behaved entirely normally for the week of pumping. The pigs were not anticoagulated.

At autopsy, no significant pathology could be found.

More complete descriptions of this experiment are found in Ref. 10.

Because the system was safely designed and proven atraumatic in animal experiments, we felt that initial clinical use of the balloon was justified.

Hemodynamic effectiveness of the balloon in human use has always been exactly what one would expect from the original animal experimentation. In Fig. 9, a pulse contour measured on the neck of a patient is seen. In this figure, the larger pulse is due to the balloon.

These hemodynamic effects have resulted in improvement in a variety of parameters of cardiovascular status. Although the results are still preliminary, blood pressure has increased, the need for supportive drug therapy has decreased, urine output and ectopic activity have improved when patients were treated with the balloon. Coupled with this improvement is the fact that no significant pathology has resulted from the use of the balloon in these patients, and no harmful after effects from therapy have been found.

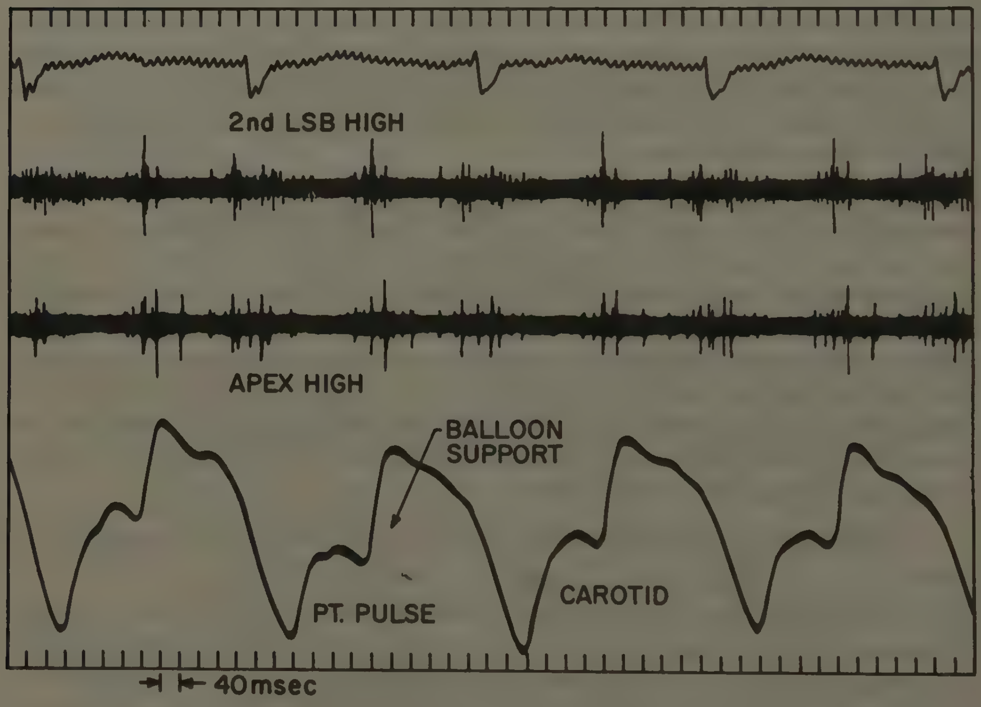


Fig. 9. Carotid pulse tracing of a patient in whom the balloon was used. Note the large diastolic pulse due to the balloon.

REFERENCES

1. Kantrowitz, A. and Kantrowitz, A., Experimental Augmentation of Coronary Flow by Retardation of the Arterial Pulse. *Surgery,* **34,** 678, 1953.
2. Clauss, R. H., Birtwell, W. C., Albertal, G., Lunzer, S., Taylor, W. J., Fosberg, A. M., and Harken, D. E., Assisted Circulation. I. The Arterial Counterpulsation. *J. Thoracic & Cardiovascular Surg.,* **41,** 447, 1961.
3. Birtwell, W. C., Soroff, H. S., Giron, P., Thrower, W. B., Ruiz, U., Deterline, R. A., Synchronous Assisted Circulation, *Canad. Med. Assoc. J.,* **95,** 652, Sept. 24, 1966.
4. Kantrowitz, A., Akutsu, T., Chaptal, P. A., Krakauer, J., Kantrowitz, A. R., Jones, R. T., a clinical Experience with an Implanted Mechanical Auxiliary Ventricle, J.A.M.A. **197,** 525, August 15, 1966.
5. Moulopoulos, S. D., Topaz, S., Kolff, W. J., Extracoporeal Assistance to the Circulation and Intra-Aortic Balloon Pumping Teams, Amer. Soc. Artif. Int. Organs, VIII:85-89 (1962). (See also Am. Heart J., **63,** 669-675 (1962).
6. Sarnoff, S. J., Braunwald, E., Welch, G. H., Case, R. B., Stainsby, W. N., Marcruz, R., Hemodynamic Determinants of Oxygen Consumption of the Heart with Special Reference to the Tension-Time Index. *Am. J. Physiol.* **192,** 148, 1958.
7. Jones, R. T., Petschek, H. E., Kantrowitz, A. R., Elementary Theory of Synchronous Arterio-Arterial Blood Pumps, *Med. & Biol. Engrng.,* **6,** 303 (1968).
8. Laird, J. D., Madras, P. N., Jones, R. T., Kantrowitz, A. R., Kothari, M. L., Buckley, M. J., Austen, W. G., Theoretical and Experimental Analysis of the Intra-Aortic Balloon Pump. Trans. Amer. Soc. Artif. Int. Organs, Vol. XIV, 1968.
9. Hallock, P., Benson, I. C., Studies on the Elastic Properties of Human Isolated Aorta, *J. Clin. Invest.* **16,** 595 (1937).
10. Madras, P. N., Laird, J. D., Iatridis, E., Kantrowitz, A. R., Buckley, M. J., Austen, W. G., Effects of Prolonged Intra-Aortic Balloon Pumping, Trans. Amer. Soc. Artif. Int. Organs, Vol. XV, 1969.
11. Bruner, J. M. R., Hazards of Electrical Apparatus—A Review, Anesthesiology **28. 2.** March-April 1967.

THE SCANNING ELECTRON MICROSCOPE: MULTI-RADIATION, NONFOCUSED IMAGING

Thomas L. Hayes

PART 1

Introduction

Information transfer, along with resolution, plays a vital role in the imaging of biological systems. Resolution might be defined as the minimum volume of the specimen that is discretely represented in the image. Information transfer describes what we learn about that volume in terms of the kinds of information transferred to the image. During the past several decades, the emphasis of work in microscopy has been directed towards higher and higher resolution. This has come about partially because of our inheritance from the light microscope. The light microscope is an imaging system rich in information, but limited in resolution. The light microscopist tended to take for granted the many aspects of information, (the location of chemical bonds, the ability to stain specifically, the ability look at living systems, the ability to look at systems in three dimensions.) as part of the inherent nature of all microscope systems. He was, however, painfully aware of the lack of resolution; he could not see very small objects. In particular, he could not visualize the viruses, macromolecules and many cell organelles. The light microscope was limited in resolution, not because of an imperfect manufacture of the instrument, but because of certain physical principles that dictate the limit to the resolving power of a lens optical system. The image formed in a light microscope is essentially limited to a resolution of approximately one half of the wavelength of the radiation being used, and in the case of visible light this would result in a resolving power of about two thousand angstroms. The direction of research, then, was towards an instrument that would provide higher resolving power. To do this the only parameter available was the wavelength of the radiation being used. While ultraviolet

---

This work has been supported in part by the United States Atomic Energy Commission.

radiation can be used, the increase in resolving power is not large—and if we go to higher frequency electromagnetic radiation, that is x-ray, or gamma ray, the lens systems necessary for focusing are not available. It was thus necessary to turn to the dual nature of particles and take advantage of the fact that not only is electromagnetic radiation characterized by a dual nature, sometimes being looked at as a wave and sometimes as a particle, but also all particles have a dual nature and can be described usefully at times as a wave. If we examine the wavelength of an electron beam of reasonable energies we find that the wavelength is very short indeed. In addition, lenses are available for electrons and a highly developed field of electron optics exists. Thus the electron microscope uses the same imaging techniques (spatial focusing of the radiation leaving the specimen) to form an image, but does so using a very short wavelength radiation. This results in high resolution and we are familiar with the standard transmission electron microscope operating with a resolution of less than 5 angstroms. While we have obtained this very high resolution, we have not paid so much attention to what has happened to the information content of the images formed—and, as it turns out, the transmission electron microscope image, while being very high in resolution is somewhat limited in information content. First the electron radiation interacts with the specimen at an atomic level rather than at a molecular level, that is the electron beam does not recognize chemical bonds but is dependent to a first approximation only on the atomic number and density of the material. This results in the loss of the ability to recognize chemical entities in terms of the chemical bonds present which was so powerful a tool both for determining naturally occurring chemical makeup and for using stains in the light microscope. Secondly, the electrons will not penetrate a very thick specimen. In fact specimens of about 1/50th of a micron are required for the usual transmission electron microscope preparation. This means that we are limited to a two dimensional image and this image again lacks some of the information that could be obtained, for example, with a stereoscopic light microscope. Since the specimen must be placed in a vacuum in a transmission electron microscope, and since it is necessary to section the specimen, the possibility of looking at living material is lost.

Our interest is to examine ways in which we could increase the information content of the image, as well as the resolving power. The scanning electron microscope (SEM) is a system which provides a high information content image at resolutions that are well above the light microscope but not as high a resolution as that available in the transmission electron microscope. Present day commercial instruments have a resolving power in the range of one hundred to two hundred angstroms. . . one order of magnitude less than the resolution that can be obtained with the transmission electron microscope. The SEM does, however, offer the possibility of utilizing a variety of modes of operation, each using a particular kind of radiation to carry the information about the specimen point being investigated. The scanning electron microscope forms an image using a principle that is different from the light microscope and the transmission electron microscope. Instead of utilizing a lens for spatial focusing in order to create the one to one correspondence between points on the specimen and points on the image, the scanning electron microscope utilizes a time sequence of points to form this one to one correspondence (Fig 1). Thus it is very similar to the image forming system used in television where a twin beam system is linked together by a synchronizing pulse so that both beams sweep over their raster in synchrony and at each instant there is an information pulse or a video signal that determines the information of the image point on the display beam by modulating that beam according to the intensity of the video signal. In the light microscope and the transmission electron microscope, the two functions of the microscope, localization and information transfer, are carried out by the same radiation. In the light microscope, visible light is focused to a point and this same radiation interacts with the specimen to form the image characteristics. In the electron microscope a similar lens system is used. Electron radiation interacts with the specimen and this same radiation is used to form the image through a lens. In the scanning electron microscope, the two functions of localization and information transfer are uncoupled, that is, one may be performed by one type of radiation and the second by another. For example, we might localize by electrons and carry the information by cathodoluminescence or the visible light signal that is emitted from the specimen due to electron bombardment. This signal is not focused—it does not fall in the realm of the imaging rules stated previously. Thus we may have a way to form a high information content image at high resolution. The light microscope produces an high information content image but is limited in resolution. The transmission electron microscope produces an image high in resolution but limited in information. The scanning electron microscope can act to incorporate both high resolution and high information content. It can use a multiradiation, high information transfer system and at the same time can localize at resolutions determined by electron optics.

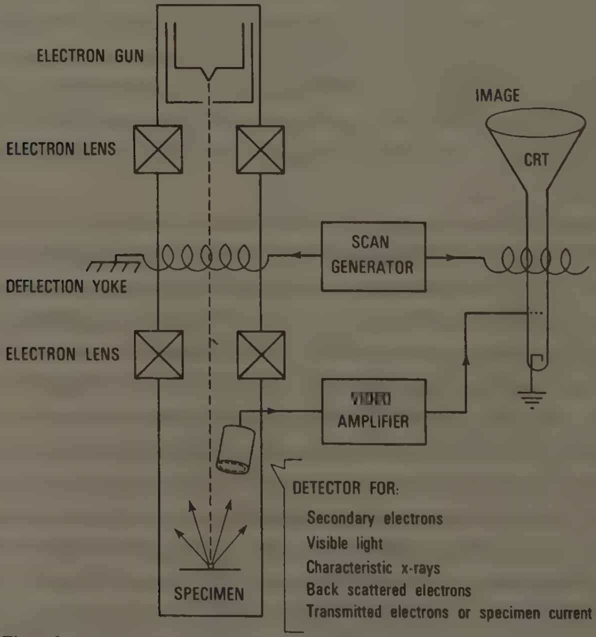

ELECTRON GUN
IMAGE
ELECTRON LENS
CRT
SCAN GENERATOR
DEFLECTION YOKE
ELECTRON LENS
VIDEO AMPLIFIER
DETECTOR FOR:
Secondary electrons
Visible light
Characteristic x-rays
Back scattered electrons
Transmitted electrons or specimen current
SPECIMEN

Fig. 1. Scanning electron microscope schematic diagram.

There is a second information transfer step of importance in the microscopy of biological systems. This transfer is from the image to the observer. Most often we consider the transfer of information from a microscope image to the observer to be carried out in the traditional analytic modes of scientific investigation. We would like to investigate the possibilities of adding to the analytic mode the subjective, intuitive or experiential modes that can be of considerable value in understanding complex systems. As a musical analog we might consider the transfer of information about a Beethoven symphony. We would suggest that all possible analytic techniques be carried out to write the program notes or other analytic descriptions of the symphony but we would also suggest that it would be useful to listen to the symphony. We are then interested in two information transfer steps; the first from the specimen to the image and the second from the image to the observer. The direction of the research is not so much toward very high resolution at whatever information content is possible but rather toward very high information at whatever resolution is possible.

PART 2

Resolution and Information Content of Electron Microscope Images

In order to illustrate the two components of a microscope image, resolution and information content, let us look at an image at the same resolution but at different information content. Fig 2 shows a wire grid that contains 200 lines to the inch and is used in transmission electron microscopy. Figure 2a is a transmission electron microscope image of this grid. Figure 2b shows the same grid at the same magification as seen in the scanning electron microscope. It is useful in our understanding of this system to have the image provided by the scanning electron microscope. It is not that the resolution is any higher but that the information content of the image is different. By looking at the scanning electron microscope picture we can not only determine measurable quantities such as the distance between the wires, but we can also observe qualities of the system that are difficult to transfer by digitizing the information into words or numbers. Wovenness might be an example of such a quality. Wovenness does not depend on a particular spacing of the wires, or on a particular angle between them, or even on the fact that there be an alternating up and down arrangement. There might, for example, be one warp and two woof lines. Wovenness is a quality that is difficult to describe in words and if we attempt to do this we often find ourselves relying on our fingers to demonstrate what we are talking about. Wovenness, then, is a quality that is transferred synthetically rather than analytically from an image to an observer. When we look at the image of the grid in the scanning electron microscope, we are immediately aware that it is a woven system. It would be rather difficult to program a computer, however, to recognize in general whether or not a system is woven.

The image of the grid (Fig 2) not only demonstrates the enriched information transfer from specimen to image, but it can also be used as an example of the

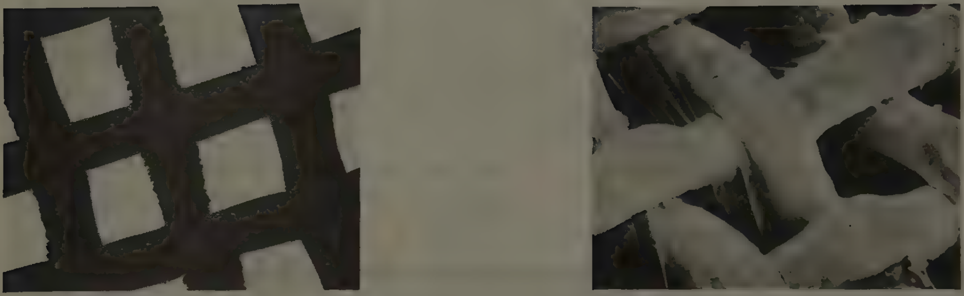


Fig. 2. 200 mesh wire grid.

2a. Transmission electron microscope. 2b. Scanning electron microscope.

possibility for synthetic, or intuitive or experiential transfer of information from the image to the observer. There are many systems in biology that are very complex in their spatial arrangements. As the system becomes more complex the value of subjective information transfer increases. In considering the appropriateness of a microscope image, therefore, both information content and resolution should be considered.

PART 3

Principles and Operation of the Scanning Electron Microscope

The scanning electron microscope might better be named a TV system that magnifies rather than an electron microscope that scans. The light microscope and transmission electron microscope operate on the same basic principle of image formation: spatial focusing through a lens to accomplish the one to one correspondence between points on the specimen and points on the image. The scanning electron microscope uses a different physical principle of image formation (Fig. 1). It utilizes a time sequence of points rather than a spatial focusing to form the image. It is a twin beam system—both beams sweeping in a raster or array of lines. The beams are synchronous—that is, they each start at the upper left corner of the raster—sweep through an identical number of lines and finish their raster at the same instant. This means that at any instant the probing beam would be at a single point on the specimen and the display beam would be at a corresponding single point on the face of the cathode ray tube. The information about this point on the specimen is transferred to the image by any one of several types of radiation that result from the bombardment of the point by the primary electron beam. As the primary electron beam strikes the specimen, it produces a variety of electromagnetic radiations (x-rays, visible light, infrared, etc.) some participate radiation (secondary electron, backscatter electron) and certain alterations in the specimen itself that can be related to the electrical properties of the speciman (induced specimen current). Any one of these radiations or current signals can be used as the information transfer signal that is used to modulate the brightness of the beam of the cathode ray tube at the corresponding point in the display raster. This means that we have available to us several modes of operation each utilizing a particular kind of secondary radiation, each of these radiations carrying its own kind of information about the specimen point under consideration.

PART 4

Information Transfer from Specimen to Image

Any one of the several modes of operation of the scanning electron microscope can select a particular kind of radiation to be used to transfer the information about the specimen. An example of this is the operation of the scanning electron microscope in the cathodoluminescent mode. The visible light produced by the bombardment of the primary electron beam is measured by a photomultiplier detector and the signal used as the video modulating signal. In this mode we have some of the information associated with visible light radiation visualized at resolutions determined by electron optics. We can also mix two or more of the modes, for example, cathodoluminescent and secondary electron modes. The secondary electron mode carries information about the topography of the surface of the specimen and the cathodoluminescent mode would demonstrate where material giving off visible light is located. If we use color coding, both kinds of information can be seen on the same image. Thus the several signals available in the scanning electron microscope can be individually coded in color and displayed in a single image as we are observing the specimen.

Another way of increasing the information transfer from specimen to image would be to use a variety of instruments, each specifically designed to utilize a particular form of radiation, as well as utilizing a scanning electron microscope in various radiation modes. This multi-instrument approach is demonstrated in a series of figures dealing with the small marine animal, the rotifer. The animal is about a tenth of a millimeter in length. Towards the center of the head area there is a jaw which represents a primitive kind of hard tissue. It is this jaw, in particular, that we will be concerned with in imaging this small animal. Fig. 3a* uses the phase con-

---

* Figs 3a, c courtesy of Dr. J. K. Koehler, Univ. of Washington, Seattle.

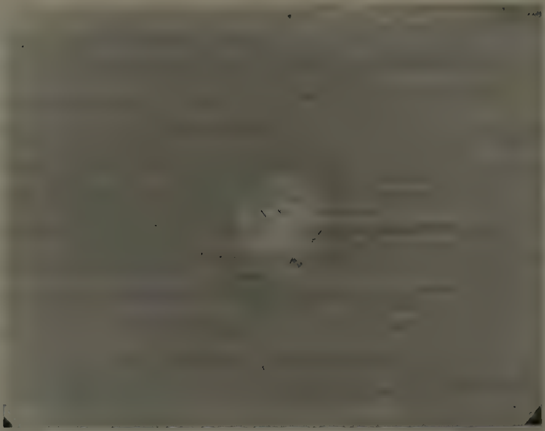
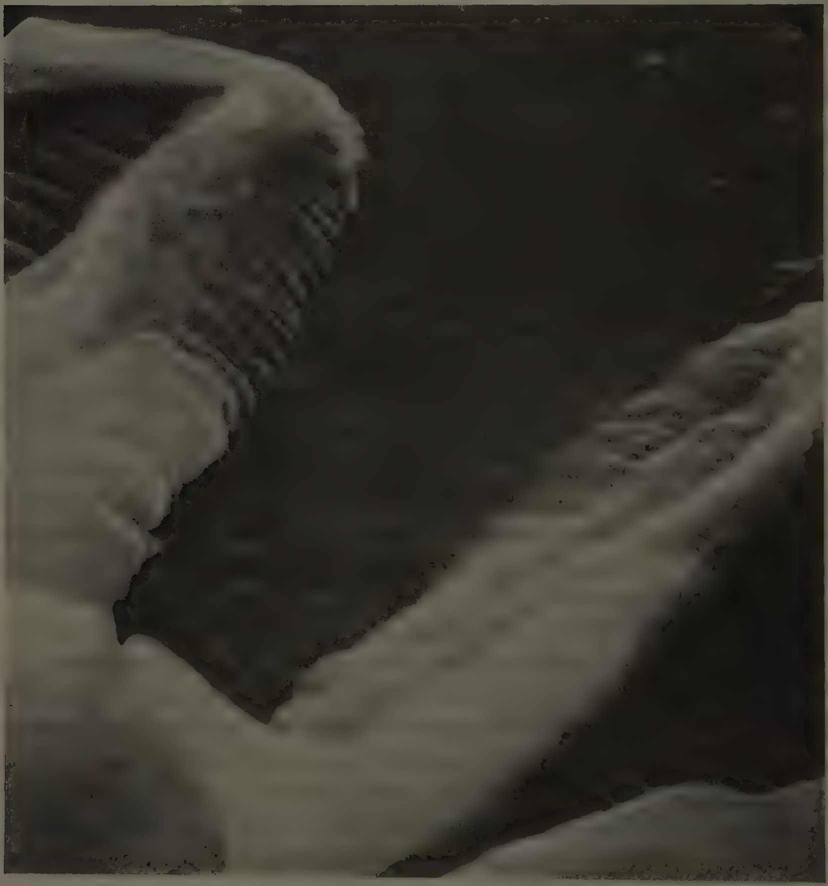
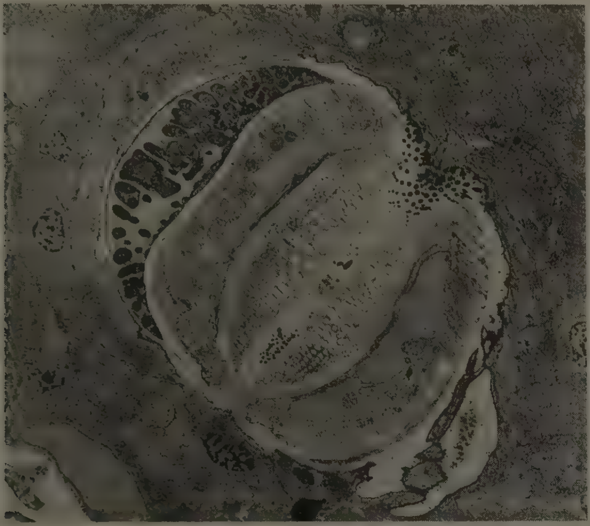


Fig. 3. Rotifer jaw.
3a. Phase contrast light microscopy. (J. K. Koehler, University of Washington, Seattle.)



3b. Scanning electron microscope.



3c. Transmission electron microscope. (J. K. Koehler, University of Washington, Seattle.)

3d. Jaw of predator species.

trast light microscope to show the structure of the isolated jaw which has been removed from the rotifer by a form of chemical dissection. We can see some of the internal structure of the jaw, but are limited in terms of our resolution.

We move then to the scanning electron microscope, and Fig. 3b shows the jaw of the rotifer as seen in the scanning electron microscope. We now find a higher resolution but still maintain three dimensional characterization of the specimen. Fig. 3c shows a transmission electron microscope section through the same jaw. The three instruments, the scanning electron microscope and transmission electron microscope and the light microscope can be used to complement one another. . . the transmission electron microscope giving higher resolution and interior structure, the scanning electron microscope adding in the interpretation of the thin section material, and the light microscope providing information about the living specimen and histochemistry. It is sometimes difficult to interpret the structures seen in thin section even with the difficult and laborious techniques of serial sectioning. And the SEM can serve as a bridge between the transmission electron microscope and the familiar light microscope image.

The jaw that we have been looking at belongs to a rotifer species that is characterized as a grazing species. Its jaw is used for grinding of food that has drifted into the mouth cavity. There is another species of rotifer that is characterized as a predatory rotifer and the jaw from this species is shown in figure 3d. This jaw is used to reach out and grab the prey and its structure conveys to many people the predatory nature of its function. It would be very difficult to program a computer to recognize a predatory jaw versus a grazing animal's jaw. It really requires the synthetic, intuitive transfer that is only possible when the image that we are viewing corresponds to experienced vision in many of its parameters.

By utilizing both the multi-radiation capabilities of the scanning electron microscope and intermicroscopy correlation possible with this instrument the amount of information transferred from specimen to image can be considerably increased.

PART 5

Information Transfer from Image to Observer

The traditional mode of information transfer from image to observer is, in scientific work, the analytic, objective mode. Information in general is reduced to words, numbers or ideas and the sum total of the image is described in these digitized bits of information. We would like to consider adding subjective, experiential modes of information transfer to the traditional ways of looking at an image. The value of such subjective information transfer depends somewhat on the definition of reality that we prefer. Many may not agree with Soren Kierkegaard when he says that existence precedes essence but the influence of his thinking is an existential fact and we feel that at least the possibilities of experiential information transfer should be investigated in order that we can utilize more of the person and, in particular, utilize the intuitive power of the individual. If we limit ourselves to what is the same for all of us, that is, the classical physics, chemistry, mathematics, biology, we may not be using the best of each of us. Rather than only analyze the image as is the classical scientific approach, we would like to add the synthetic or artistic approach to our ways of understanding the image. The scientist, in general, fits the problem to his tools and arrives at a very neat solution but at times has so reduced the problem that it is no longer of any importance. The artist, on the other hand, attempts to fit his tools to the problem and while his solution is not neat and varies from person to person, he can always say that he is still dealing with the real problem.

The choice of a definition of reality is, however, an option to be exercised by the individual. Some may agree, for example, with Plato that there is an essence of treeness that is greater than any individual tree in that each individual tree merely approximates this essence. On the other hand, we could choose in favor of the point of view of Kierkegaard, who says that this essence of treeness is merely the lowest common denominator of all the characteristics of the trees and that each individual tree has this essence and also qualities specific to its own existence. Without trying to decide this particular issue, however, we feel it is at least important to investigate the effect of adding subjective, intuitive information transfer to our microscope study of biological systems. In order to experience the image, it is necessary that the image contain the parameters of experienced vision—not everything that comes through the eyes is experiencd vision and possibly the most extreme example of this is reading. In reading, the information comes through the eyes but this information does not carry with it the confidence that we associate with experienced vision.

Often times the scientist will turn to the artist to take data that he has abstracted from the system and put it back into reality. An example of this was the movie, Fantastic Voyage, in which a group of individuals were shrunken into a very tiny submarine and floated through the blood stream. This picture was made by having an artist take information that had been accumulated by scientific analysis from a group that we might broadly classify as biophysicists and the artists put the information back into reality by adding parameters of experienced vision. One difficulty with this process is that the information given to the artists has already been abstracted—it has been arrived at through the process of data reduction. Such a process means that the scientists has already chosen which parts of the system are important or suitable for his study and therefore it is only this limited part of the system that is converted into an experience for the observer. What we would like to do is to eliminate the artist and the biophysicist and produce images that correspond to experienced vision directly; to produce a microscope that sees in the same way that our brain sees. We cannot shrink ourselves into submarines and float through the blood stream but we can attempt to determine the parameters of experienced vision and to extract directly from the system these same parameters.

One important parameter is the number of dimensions appropriate to our image. In the high resolution light microscope, and in the transmission electron microscope the image is essentially a two dimensional image. Now we often appreciate the fact that we are not four dimensional creatures, that we do not understand four dimensions in the same way that we can understand or experience three dimensions. We do not so often remind ourselves that we are also not two dimensional creatures—we are three dimensional donkeys not two or four or *n* dimensional creatures. It would then be important to experiential transfer if we could add to our image information having to do with depth in the sample that we are studying. The brain sees depth by a variety of techniques. Possibly the most used code for depth information is learned size, that is, the brain compares the size of the image of the retina with the size of a stored image and calculates from this ratio how far away the object must be. A second way that we learn about depth is through binocular vision utilizing parallax. The brain takes the images from both eyes, compares the shift of the object being viewed against a distant background as viewed in the left and right eye and calculates from this shift how far away the object must be. A third code for depth and the one used in producing scanning electron microscope images has to do with the radiation intensity leaving each point on the subject being viewed. The brain takes the intensity of visible light leaving the surface of an object, calculates the intensity and then integrates this intensity over the surface to give shape. This is a rather familiar way that the artist has utilized to represent or to code for depth on a two dimensional canvas and is generally described as shading. A somewhat more sophisticated way has to do with pattern recognition and overlap. The overlap code is an important method of depth perception and if it is lacking, some rather bizarre effects can be seen.

If we now turn to one of the modes of operation of the scanning electron microscope, called secondary electron, we find that the function that determines the number of secondary electrons leaving the surface at each point is a function that is quite similar to the function that determines one of the depth coding devices, that is, radiation intensity as a function of the angle between line of sight and the surface. Secondary electrons leave the surface as a function of the angle between the beam and the surface. Thus, if we modulate the display cathode ray tube according to the secondary electron signal we can transfer depth information.

The next series of figures show a human blood clot and the attempt here is to experience the clot as well as to provide analytic information. We wish to see if we can make an image that appears as if the cells were actually large so that we can walk among them. Figure 4a, b, c, show increasing magnification of this human blood clot and as we increase the magnification we can perceive the cells and finally even the fibrin strands holding the cells in the clot. By utilizing a scanning electron microscope we produced a very high efficiency representation of surface. If we rely on sectioning of such a surface, the only information is transferred at the intersection of two planes—the plane of the specimen surface and the plane of the section intersect in a line. This line is a very low efficiency representation of surface information. The scanning electron microscope, however, can provide information on surfaces that is much more readily appreciated.

The interaction of morphological units at a size level that lies above the molecular level can be of considerable importance. For example, we have observed the structure of collagen in a healing wound. Mechanical engineering tests on heal-

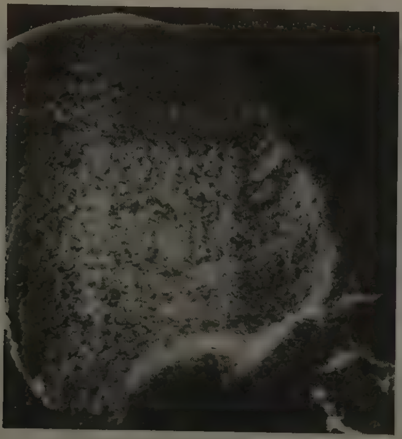


Fig. 4a

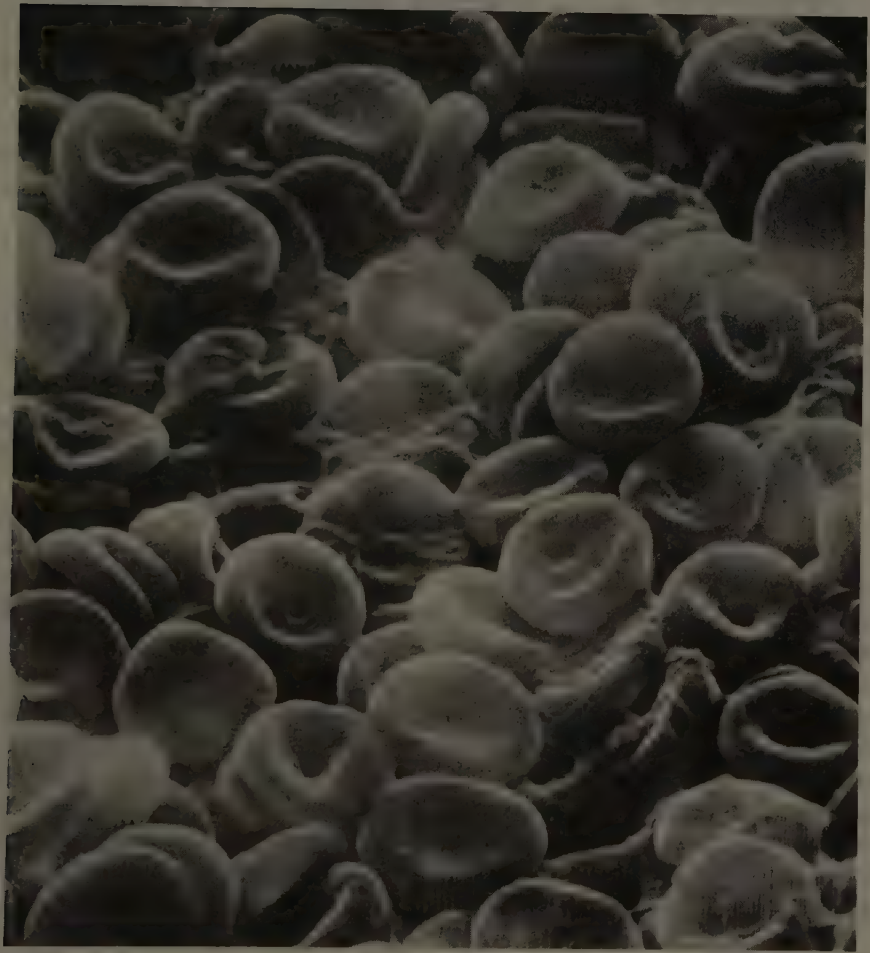


Fig. 4b

Fig. 4c

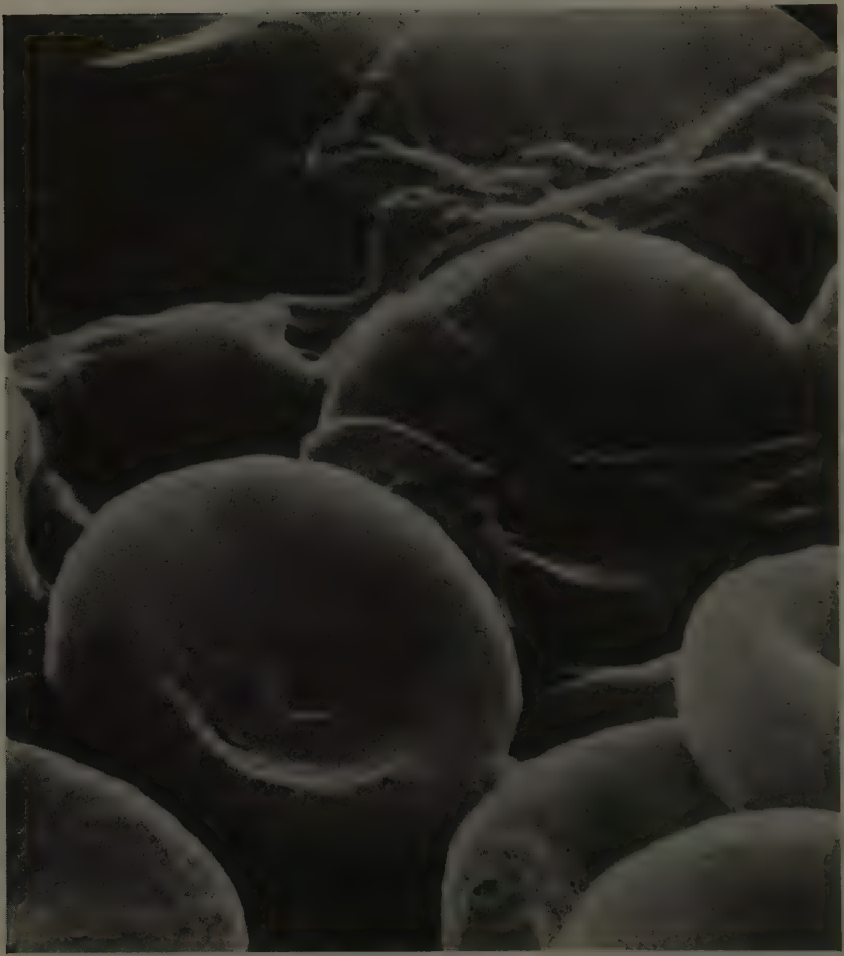


Fig. 4. Human blood clot. Taken at magnification from 100 X to 10,000 X.

ing wounds have shown that the wound does not recover its original strength, but preliminary biochemical tests of the collagen molecule seem to indicate that the molecular structure itself was intact. It was felt then that the weakness of the collagen matrix in the healing skin might be a reflection of a higher level disorganization rather than strictly a molecular phenomenon. Normal skin is made of collagen fibrils that are in a sense spun into collagen fibers—the collagen fibers in turn are woven into the tissue of the skin. It appears that in the healing wound, the ability to spin the fibrils into a fiber has been lost and we are faced with a situation that would compare to an attempt to weave cloth without first spinning thread. If we take cotton fibrils and try and weave with the cotton fibrils without the spinning wheel, we find that we would be in considerable difficulty. It would appear that in the healing wound, the inability to form fibers has resulted in a fabric that has considerably less strength than the normal skin.

Because no sectioning is necessary the same specimen on its original standard light microscope slide can be viewed in the scanning electron microscope. In addition to experiencing the three dimensional qualities of the image, we can utilize the higher resolution of the electron optical system, making it possible to observe the surface that lies at right angles to the plane of the section.

Human chromosome preparations have been observed in the scanning electron microscope. The resolution possible here is still limited by preparative techniques and is not much better than that possible with the light microscope. In the light microscope we are observing the location of stain and therefore are looking through some of the transparent protein material that overlies the chromosomes. In the secondary electron mode of the scanning electron microscope we are observing matter of any type and thus we see a covering over the chromosome. However, there is an aspect of the scanning electron microscope which may prove useful with respect to problems such as chromosome determination (karyotyping) and this is that the scanning electron microscope is its own flying spot scanner. Instead of having to form an image on a glass plate, as is done in the light microscope for karyotyping, and then placing this image in a flying spot scanner to convert the information to a digitized electronic pulse, the scanning electron microscope video signal is already such an electronic digitized pulse and can be fed directly into a computer for on-line analysis of chromosome number and aberrations.

PART 6

Observation of Living Specimens

The observation of living specimens in the scanning electron microscope has the advantage of minimum preparative techniques and minimum artifact being present. There is also the possibility of coding the observer for time, for it has been said that time is defined in such a way as to make motion simple. If motion can be observed we might be able to experience the time quality of the image as well as the three dimensional qualities spoken of in the last section. It is also possible to do physiological experiments if the specimen is living. Such experiments as micro-dissection, micro-irradiation, or even electro-physiological experiments are possible. However, the scanning electron microscope operates in a vacuum so the specimen chosen must be one that can survive the environmental conditions of vacuum, irradiation, heat, etc., present in the scanning electron microscope.

Several systems have been found that can survive for periods up to one hour in these environmental conditions. One of the most useful has been the Tribolium confusum flour beetle. We can thus, in certain circumstances, observe systems that are living (in the case of the beetle, viability was checked by the ability to proceed through metamorphosis or the ability in the case of the adult to resume normal activity). Other systems that have been studied are systems that, in contrast to the Tribolium, which conserves its water, are systems which can lose their water and still remain viable. An example of this is the cryptobiotic cytoplasm of the brine shrimp embryo.

While Tribolium and certain other species can survive in the vacuum environment they are, in general, in an anesthetized, non-moving condition. After being removed from the microscope column, it has in general taken about one hour to regain consciousness and normal movement. Thus to date using the specimens that have been shown to be viable in the microscope, no motion has actually been observed while the specimen was in the microscope column.

PART 7

Signal Processing Prior to Display

The video signal can be processed before being used to modulate the display cathode ray tube. Examples of this processing have been described (see for example, MacDonald in EMSA Meeting, 1968) and such operations as noise reduction have been carried out. The derivative of the signal has been used as the modulating quantity for the display. Using the derivative of the signal, that is, the time rate of change of intensity as the video signal, the range between black and white has been brought much closer together so that the entire image can be seen without areas that are totally black or whited out and also small differences in gray scale can be accentuated due to their rapid change of intensity rather than displaying them as the gray scale itself.

PART 8

Types of Modulation of the Display Cathode Ray Tube

In general, the video signal is used to modulate the brightness of the cathode ray tube at each instant as the beam sweeps over the face of the tube. However other possibilities are open in terms of display modulation. It might be useful to convert the intensity modulated image into a spatial modulated image. To do this, Dr. Everhart has devised a technique for deflection modulation. The beam, instead of being turned on and off according to the video signal, is deflected in the *y* direction according to this signal. The same information shown previously in intensity modulation now is represented as a spatial quantity—areas that were gray are now rough. In terms of the transfer of information from the image to the observer and in his intuitive or experiential response, the word and the concept of gray is a rather blah kind of concept. Rough, on the other hand is quite a different concept, it's rough like a file. We are interested in learning how we get ideas about systems and how we can stimulate our intuitive processes by converting the information into an experiential image.

Another possibility in terms of modulation would be to modulate the color of the image display point rather than its intensity or its *y* position. Color modulation has the advantage of retaining the one to one correspondence between points on the specimen and points on the image which deflection modulation does not do.

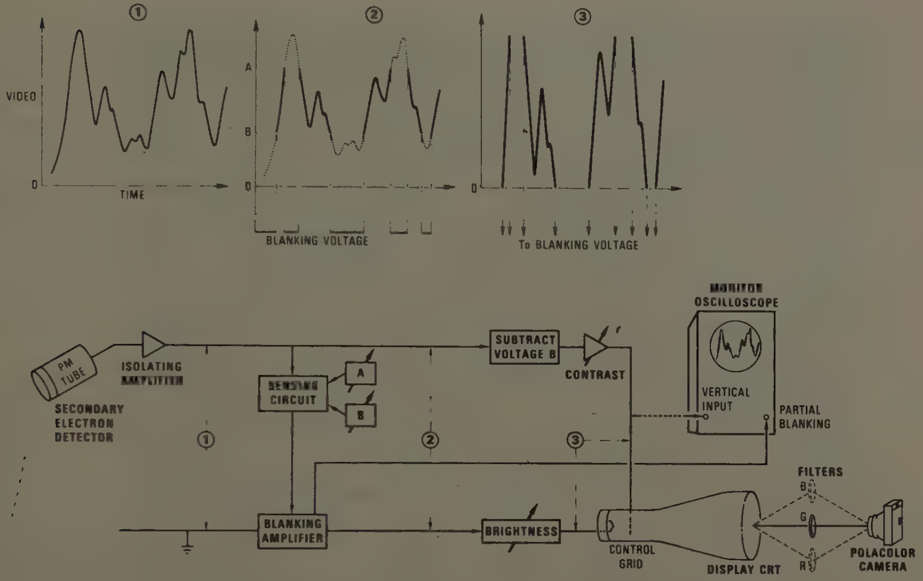

①
②
③
VIDEO
A
B
0
TIME
BLANKING VOLTAGE
To BLANKING VOLTAGE
MONITOR
OSCILLOSCOPE
PM TUBE
ISOLATING
AMPLIFIER
SECONDARY
ELECTRON
DETECTOR
SENSING
CIRCUIT
SUBTRACT
VOLTAGE B
CONTRAST
VERTICAL
INPUT
PARTIAL
BLANKING
FILTERS
B
G
R
BLANKING
AMPLIFIER
BRIGHTNESS
CONTROL
GRID
DISPLAY CRT
POLACOLOR
CAMERA

Fig. 5. Color modulation schematic diagram.

Figure 5 shows a schematic diagram of the color modulation circuitry used. The video signal passing through a window that allows only a selected intensity range of the signal to be displayed on the cathode ray tube. Each intensity then, through multiple exposure, using the light from the standard cathode ray tube, is filtered and coded in a particular color. Thus, instead of intensity modulation we have a color modulation possibility. We are interested not only in the ability to perceive small differences in intensity because of a radical change in color, coded for those small differences, but we are also interested in the subjective or intuitive response to the images that are painted in a variety of color. This method allows us a fairly rapid way to change the color composition of the image and through this to investigate the possibility of color as a trigger mechanism for ideas and understanding of the system.

Finally, I would like to refer to a "Peanuts" strip by Mr. Charles Schulz, which shows in a concise way what has taken a rather long time to relate in this paper. As we look at the strip, Peppermint Patty at the beginning says, "I like numbers, numbers are beautiful". She goes on to say how numbers have qualities that refer to subjective characteristics—twos are nice, threes are mean and so forth. The boy ahead of her in school turns around and says, "have you done those division problems for tomorrow?" and Peppermint Patty in the final frame says, "Nothing spoils numbers faster than a lot of arithmetic". We might suggest, to paraphrase Peppermint Patty, that nothing spoils an image like a lot of analysis—but actually we do not wish to disparage the analytic techniques. Let us analyze all that we can but leave ourselves open to the addition of subjective, intuitive processes in the transfer of information about the biological system.

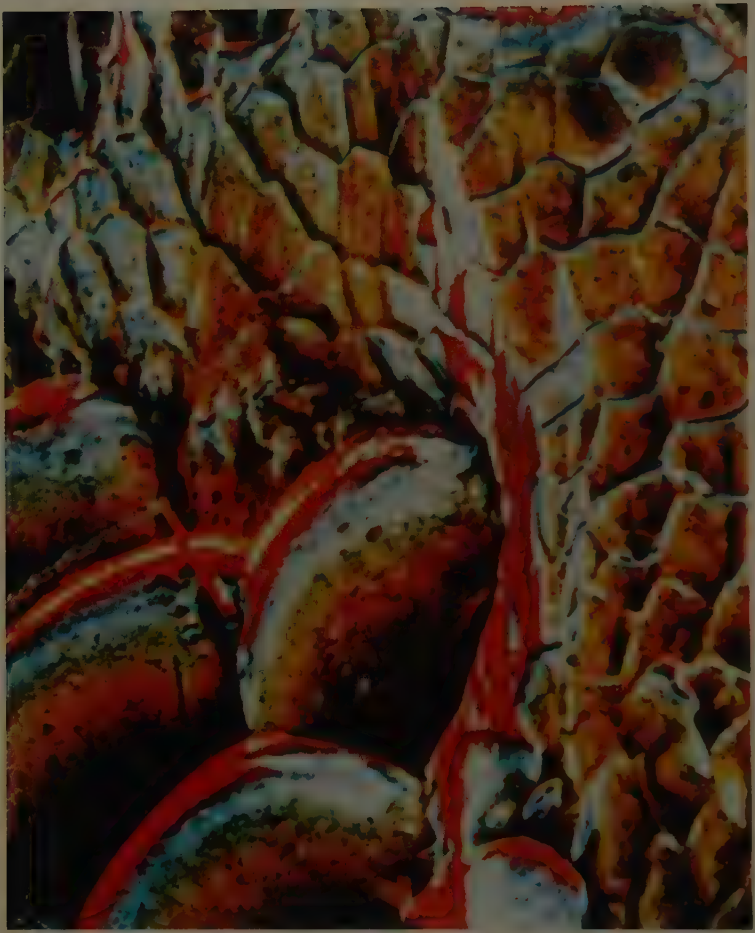


Fig. 6. Tribolium eye, color modulation, X6,000.

REFERENCES

1. Adv. in Biol. & Med. Physics Vol. 12, 1968, Academic Press.
2. Exp. Mol. Path. **10**, 186-198 (1969).
3. J. Ult. Res. **27**, 402-418 and 419-434 (1969).
4. U. C. Lawrence Rad. Lab. Report UCRL #18876 (1969).
5. Invest. Ophthalmology **7**, 651-662 (1968).
6. Nature **221**, 373-374 (1968).

NEW FORMULATION OF EARLIER RESULTS IN THE KINETIC THEORY OF TRACERS. APPLICATION TO VARIOUS PROBLEMS IN METABOLISM

Claude Kellershohn

INTRODUCTION

Since Von Hevesy (1) established the basis of the tracer method, a great deal of work has been done by many researchers on tracer dynamics and the possibilities they offer for the study of biological mechanisms. Starting from a stochastic model of the metabolic system in steady state according to Bergner (2,3,4), although in slightly different form, we shall deduce the more general relation between the flux of matter (tracee) through the system, the quantity of matter in the system and the experimental data of tracer kinetics. This general relation is independent of the number of hypothetical *states* or compartments involved by the system model and contain well known relations in use for a long time. The results obtained are applied to various examples of metabolism.

The following text which is devoted to a special aspect of tracer dynamics does not, therefore, contain due mention of all the studies that have notably contributed to knowledge in the field of interest, and the reader is referred to the fine general reviews (5-13) and treatises (14,15) and proceedings (16,17,18) that provide a comprehensive view.

REPRESENTING THE METABOLISM OF A SUBSTANCE BY A MARKOV PROCESS

Basic relation.

The particles of matter constituting the system are considered as distributed in a finite number of states n and the particles in any given state are considered as forming a *compartment*. In the limiting case, if all particles are in individual states, there will be one particle per compartment. From this standpoint, we may use the terms "state" and "compartment" without distinction. The parameters of the system shown schematically in the figure 1, i.e. Q the quantities of matter in each compartment, R the matter transfers from compartment to compartment

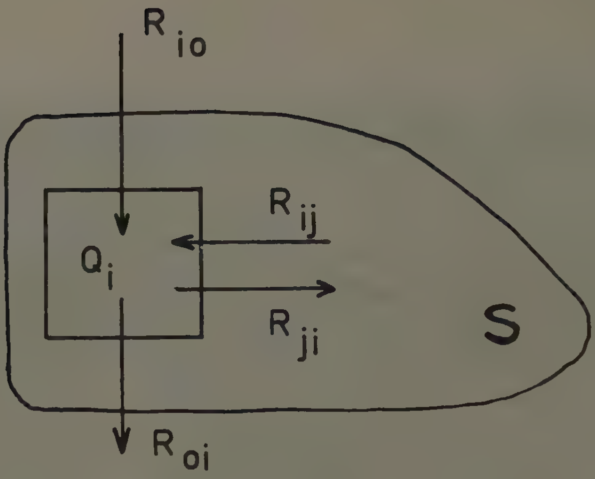


Fig. 1

or efferent transfers, and k the rates of transport from compartment to compartment, are represented by symbols as proposed by the task group on tracer kinetics of the I.C.R.U. (12):

Q_i = quantity of matter in compartment i

k_{ji} = transport rate constant in compartment j from i

$R_{ji} = k_{ji}\, Q_i$ = transport into compartment j from i

R_{io} = transport into system at compartment i.

If the parameters of the system are time functions, application of the principle of the conservation of matter produces the following relations (applied to each compartment):

$$\frac{dQ_i(t)}{dt} = R_{io}(t) + \sum_{\substack{j=1\\ j\neq i}}^{n} R_{ij}(t) - \sum_{\substack{j=1\\ j\neq i}}^{n} R_{ji}(t) - R_{oi}(t) \qquad i = 1, 2 \ldots n.$$

If the system is in the steady state $\frac{dQ_i}{dt} = O$, parameters are time independent and the foregoing relations may be expressed as follows:

$$R_{io} = - \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{R_{ij}}{Q_j} Q_j + \left[\sum_{\substack{j=1\\ j\neq i}}^{n} \frac{R_{ji}}{Q_i} + \frac{R_{oi}}{Q_i} \right] Q_i \qquad i = 1, 2 \ldots n \tag{1a}$$

it being taken that

$$k_{ij} = \frac{R_{ij}}{Q_j} \quad \text{and} \quad k_{ii} = - \left[\sum_{\substack{j=1\\ j\neq i}}^{n} \frac{R_{ji}}{Q_i} + \frac{R_{oi}}{Q_i} \right] \tag{2}$$

or:

$$R_{io} = \sum_{j=1}^{n} -k_{ij}\, Q_j \qquad i = 1, 2 \ldots n \tag{1b}$$

These n relations are expressed by the simple matrix relation

$$R = KQ \tag{1c}$$

R, K, and Q respectively designating the column vector $||R_{oi}||$, the matrix $||-k_{ij}||$, and the column vector $||Q_j||$.

Developed more than 10 years ago by BERMAN and SCHOENFELD (19), this matrix relation is quite fundamental. It represents a relation between the flow into a system in the steady state and the quantity of matter contained in it, in general form. Matrix K governs the relation and characterizes the properties of the system from the kinetic standpoint. Its elements have the dimensions of the inverse of a time and appear as a kind of generalization of the concept of the "turnover rate".

Relation (1c) may be expressed in the form

$$Q = \Theta R \tag{1d}$$

that is to say

$$Q_j = \sum_{i=1}^{n} \theta_{ji} R_{io} \tag{1e}$$

$\Theta = K^{-1}$ being the inverse matrix of K and characterizing the system, like K. Its elements have the dimension of a time and in a sense represent the concept: "mean time of sojourn", in general form. We shall see that the information that the use of tracers can provide concerning such a system are largely derived from the simple relation existing between their kinetics and the matrix Θ. We call Θ the *occupancy matrix.*

Markov process.

In the foregoing, we have implicitly considered that in the steady state such values as R and Q are perfectly determined. In fact, however, if the evolution of a particle is considered, it must be regarded as a random change from state i to stage j. Considered from this angle, R and Q are random magnitudes with a mean value and a standard deviation.

A ratio such as $\dfrac{R_{ji}\,\Delta t}{Q_i}$ or the fraction of matter moving from compartment i to compartment j in time $\triangle t$, appears in the new terms adopted, as the fraction of particles in state i which pass the state j in time Δt. This ratio is a statistical frequency corresponding to a sharply determined probability $k_{ji}\triangle t$. The "transport rate constant" k_{ji} is then a determined magnitude, unlike R_{ji} and Q_i since it corresponds to a probability per unit of time.

The system is represented as a quantity of particles in a finite number of states n. A particle in state i and at instant τ has a probability $p_{ji}(t-\tau)$ of finding itself in state j at instant $t > \tau$ which is independent of τ and dependent solely on the difference $t - \tau$ between the two instants considered. The latter property is evidently linked to the fact that the system is in the steady state. If we consider the probability of transition between instants t and $t + \triangle t$ we find that it is given by:

$$\begin{aligned} p_{ji}(\Delta t) &= k_{ji}\,\Delta t + 0(\Delta t) \\ p_{ii}(\Delta t) &= 1 + k_{ii}\,\Delta t + 0(\Delta t) \end{aligned} \qquad (3)$$

$0\ (\triangle t)$ being the conventional sign representing a function that tends towards 0 with $\triangle t$ and in an order greater than 1. The second of these relations represents the probability that a particle has of remaining in state i during a time interval $\triangle t$, i.e., the ones complement of its probability of leaving that state. Its value is derived from the second relation (2), which (2) show that:

$$\sum_{j=1}^{n} p_{ji}(\Delta t) = 1 + \sum_{j=1}^{n} k_{ji}\Delta t + 0(\Delta t) = 1 - \frac{R_{oi}}{Q_i}\,\Delta t + 0(\Delta t)$$

which leads to the two relations:

$$\sum_{j=1}^{n} p_{ji}(\Delta t) < 1 \qquad (4a)$$

$$\sum_{j=1}^{n} k_{ji} < 0 \qquad (4b)$$

The properties indicated represent the system by a "time homogenous Markov process with discrete states in continuous time" (20) and characterised by the matrix : $P(t) = ||p_{ji}(t)||$.

Owing to the inequation (4a), matrix $P(t)$ is said to be *substochastic* (21) as against *stochastic* matrices for which relation (4a) is by definition an equation. The property of homogeneity results from the steady state and the fact that the probability of transition from state i to state j between instants τ and t depends solely on $t - \tau$. Finally, the inequation (4b) signifies that the Markov process considered is *non-conservative*. This is due to the fact that the system represented by the process is considered open. In the case of a closed system, all the R_{oi} magnitudes would have value 0, inequation (4b) would be an equation, and the process representing the system would be said to be *conservative*.

The joint probability theorem gives us the well known relation :

$$p_{ji}(t + \Delta t) = p_{ji}(t)\,(1 + k_{jj}\Delta t) + \sum_{l \neq 1}^{n} k_{jl}\,p_{li}(t)\Delta t + 0(\Delta t)$$

i.e. for $\Delta t \to 0$

$$\dot{p}_{ji}(t) = \sum_{l=1}^{n} k_{jl}\,p_{li}(t) \qquad (5a)$$

or in matricial notation

$$\dot{P}(t) = -KP(t) \qquad (5b).$$

Integration of this differential matricial relation gives us:

$$P(t) = P(0)\,e^{-Kt}$$

matrix e^{-Kt} being defined by:

$$e^{-Kt} = \sum_{r=0}^{\infty} (-K)^r \frac{t^r}{r!}$$

and matrix $P(0)$ being none other than the unity I matrix according to relations (3).

One of the well known properties of matrix $-K$ that results from inequation (4b) is that all its proper values have a negative real part. (3,22)

This means that all terms $p_{ji}(t)$ of matrix $P(t)$ tend exponentially towards 0 as t increases indefinitely. Contrarily, in closed systems where inequation (4b) is replaced by an equation, $P(t)$ tends to become a limit matrix independent of time and corresponding to an equilibirum state for the Markov process.

If we now integrate the relation (5b) we have:

$$\int_0^{\infty} \dot{P}(t)dt = -K \int_0^{\infty} P(t)dt$$

$$P(\infty) - P(0) = -K \int_0^{\infty} P(t)dt.$$

Hence

$$-I = -K \int_0^{\infty} P(t)dt$$

$$\int_0^{\infty} P(t)dt = K^{-1} = \Theta \qquad \text{(6a).}$$

We see that the matrix Θ of the system is the integral taken from O to infinity of the matrix $P(t)$ characterising the Markov process that represents the system. This permits formulation of an interpretation of matrix Θ. In fact, integration in parts gives us:

$$\int_0^{\infty} - t\,\dot{P}(t)dt = \Big| -tP(t) \Big|_0^{\infty} + \int_0^{\infty} P(t)dt.$$

The system considered being assumed open, $|tP(t)|_0^{\infty}$ is nil, according to what we have shown above. We therefore have:

$$\int_0^{\infty} P(t)dt = \Theta = \int_0^{\infty} t[-\dot{P}(t)]dt \qquad \text{(6b).}$$

On the other hand, matrix $-\dot{P}(t)$ obeys the relation:

$$\int_0^{\infty} [-\dot{P}(t)]dt = -[P(\infty) - P(0)] = I$$

as matrix unity I has the same function as the number 1 in the matrixial fields, we may consider matrix $-\dot{P}(t)$ as a sort of generalisation of the probability density function. Relation (6b) then shows that matrix Θ appears as a sort of extension to the matrixial field of the concept of a mean particle evolution time during the Markov process.

In fact, if N_i is the number of particles in state i at instant 0, the fraction of this number in state j at instant t and which will have left this state at instant $t + \Delta t$ is:

$$\frac{N_i[p_{ji}(t) - p_{ji}(t + \Delta t)]}{N_i} = \frac{p_{ji}(t) - p_{ji}(t + \Delta t)}{\Delta t} \Delta t$$

or when $\Delta t \rightarrow dt$

$$- \dot{p}_{ji}(t)dt.$$

This result gives an interpretation of terms

$$\Theta_{ji} = \int_0^\infty - t\, \dot{p}_{ji}(t)dt$$

of matrix Θ. However, an important observation should be made in this respect. If j is different from i the expression $- \dot{p}_{ji}(t)dt$ does not represent a probability, as it can assume negative values. By the same token, the function $- \dot{p}_{ji}(t)$ is not a probability density function, because

$$\int_0^\infty - \dot{p}_{ji}(t)dt = -[p_{ji}(\infty) - p_{ji}(0)] = 0$$

is other than 1. The time θ_{ji} for $j \neq i$ does not represent the mean value of a time.

Contrarily, if $j = i$ the expression — $\dot{p}_{ii}(t)dt$ is always positive and the function $- \dot{p}_{ii}(t)$ is such that:

$$\int_0^\infty - \dot{p}_{ii}(t)dt = -[p_{ii}(\infty) - p_{ii}(0)] = -[0 - 1] = 1.$$

They therefore represent respectively a probability and a probability density function. θ_{ii} is that the mean value of a time: it represents the *mean time of sojourn* or *mean residence time*, or *turnover time* of the particles in state i.

Generally, the same reasoning demonstrates that if we consider a sub-set S' of the set of n states S, for example those ranging from $j = l$ to $j = m$ with $1 \leqslant l < m \leqslant n$, the time

$$\theta_{S'i} = \sum_{j=l}^{m} \theta_{ji}$$

is or is not the mean value of a time, according to whether the state i belongs or does not belong to the state S'. In the first case, the time $\theta_{S'i}$ is the mean sojourn or mean residence time, or again the mean turnover time of the set of states S', for the particles initially in the state i. This is of course always true for the time θ_{Si} since this state i belongs by definition to the set of n states S.

Integration of the relation (5b) after multiplying each number by t gives us a new expression of the matrix Θ:

$$\int_0^\infty t[-\dot{P}(t)]dt = \Theta = K\int_0^\infty tP(t)dt \qquad (7a)$$

$$\theta = \left[\int_0^\infty P(t)dt\right]^{-1} \int_0^\infty tP(t)dt \tag{7b}$$

$$\theta_{ji} = \sum_{l=1}^{n} - k_{jl} \int_0^\infty t\, P_{li}(t)dt \tag{7c}$$

Now, to take into account the fact that the terms k represent perfectly determined transition probabilities, we change the form of the second relation (2) and write:

$$k_{ii} = -\left[\sum_{\substack{l=1\\ l\neq i}}^{n} k_{ji} + k_{oi}\right]$$

Relation (7c) may be expressed as follows:

$$\theta_{ji} = \int_0^\infty t \sum_{\substack{l=1\\ l\neq j}}^{n} [k_{lj}\, p_{ji}(t) - k_{jl}\, p_{li}(t)]\, dt + \int_0^\infty t\, k_{oj}\, p_{ji}(t)\, dt.$$

Let us consider the second term of the second member, which we shall represent by:

$$t_{ji} = \int_0^\infty t\, k_{oj}\, p_{ji}(t)\, dt \tag{8a}$$

The expression $k_{oj}\ p_{ji}\ (t)\ dt$ represents the probability of a particle situated initially in state i leaving the set S of n states at the level of state j, between instants t and $t + dt$. We shall call this time t_{ji} the *"transit time"* of particles through the system from state i to state j. It will be observed that the transit time is not a mean value, since $k_{oj}\ p_{ji}\ (t)$ is not probability density function as:

$$\int_0^\infty k_{oj}\, p_{ji}(t)\, dt = k_{oj} \int_0^\infty p_{ji}(t)\, dt = k_{oj}\theta_{ji}$$

for all values of j, unless $k_{ol} = 0$ for every $l \neq j$.

Relation (7c) may be finally expressed as:

$$\theta_{ji} = \int_0^\infty t \sum_{\substack{l=1\\ l\neq j}}^{n} [k_{lj}\, p_{ji}(t) - k_{jl}\, p_{li}(t)]\, dt + t_{ji}$$

If we now summate times θ_{ji} and t_{ji} for all the states of the system, we obtain:

$$\sum_{j=1}^{n} \theta_{ji} = \theta_{Si} = \int_0^\infty t \sum_{j=1}^{n} \sum_{\substack{l=1\\ l\neq j}}^{n} [k_{lj}\, p_{ji}(t) - k_{jl}\, p_{li}(t)]\, dt + \sum_{j=1}^{n} t_{ji} \tag{8b}$$

The expression under the sign $\int$ in the first term of the second member is evidently nil and we have:

$$\theta_{Si} = \sum_{j=1}^{n} t_{ji} = t_{Si}$$

θ_{Si} is effectively the mean of a time as earlier, the same being therefore true for t_{Si}. We will therefore say that the mean time of sojourn or mean residence time, or turnover time in system S of particles initially in the state i is equal to the mean transit time in the system of the same particles.

EXPERIMENTAL DETERMINATION BY TRACER OF THE MARKOV PROCESS REPRESENTING A METABOLIC SYSTEM

Let us assume that we have radioactivity q_{io} in compartment i precisely at instant 0 (impulse injection). The result will be radioactivity $q_j(t)$ in compartment j at instant t. The ratio $\frac{q_j(t)}{q_{io}}$ represents the fraction of particles of tracer initially in compartment i, and which are in compartment j at instant t. The quantities $q_j(t)$ and q_{io} and their ratios are random magnitudes, as R and Q relative to the matter. We consider the tracer a perfect tracer if the mean value of the ratio $\frac{q_j(t)}{q_{io}}$ is equal to the probability of transition $p_{ji}(t)$ of the Markov process representing the metabolic system. This is in agreement with the habitual concept, which may be reduced to the statement that if a tracer is perfect it will be obey the differential relation:

$$\frac{q_j(t)}{q_{io}} = \sum_{\substack{l=1 \\ l\neq j}}^{n} \frac{R_{jl}}{Q_l}\frac{q_l(t)}{q_{io}} - \left[\sum_{\substack{l=1 \\ l\neq j}}^{n} \frac{R_{lj}}{Q_j} + \frac{R_{oj}}{Q_j}\right]\frac{q_j(t)}{q_{io}} \tag{9a}$$

which corresponds to relation (5a) which we may write:

$$\dot{p}_{ji}(t) = \sum_{\substack{l=1 \\ l\neq j}}^{n} k_{jl}\, p_{li}(t) - \left[\sum_{\substack{l=1 \\ l\neq j}}^{n} k_{lj} + k_{oj}\right] p_{ji}(t) \tag{9b}$$

The approximation of $p_{ji}(t)$ by the tracer will be the better (for statistical reasons) the greater the number of atoms of tracer used, i.e. the higher the radioactivity q_{io} injected into compartment i. Yet the number of tracer atoms must be much smaller than the number of atoms of matter in compartment i, so as not to disturb the steady state of the system. This means that (as it is generally known) the specific activity of the tracer must be as high as possible.

This being so, the experimental data supplied by the tracer give us the occupancy matrix of the system to within the statistical precision obtained:

$$\Theta = \int_0^\infty P(t)dt = \left\| \int_0^\infty \frac{q_j(t)}{q_{io}}\, dt \right\| \tag{10a}$$

with

$$\theta_{ji} = \int_0^\infty \frac{q_j(t)}{q_{io}}\, dt \tag{10b}$$

Theoretically, and as it has been recognised for quite some time (23), the n values of $q_j(t)$ ($j = 1, 2 \ldots n$) must be measured for each of the n values of

q_{io} $(i = 1, 2 \ldots n)$, i.e. that is to say that n^2 measurements are needed for complete determination of the Markov process representing the most general system of n compartments or states.

An important observation must be made at his point: for $\dfrac{q_j(t)}{q_{io}}$ to represent p_{ji} (t) it is necessary for q_{io} to be supplied as an impulse injection.

This restriction no longer applies in the case of the integral of $\dfrac{q_j(t)}{q_{io}}$ representing the element of the matrix Θ. The rate of the injection is of no importance: it is sufficient that its duration be limited. In fact, if $r_{io}(t)$ is the rate of tracer injection into compartment i, assumed to be a function of time, functions $q_j(t)$ obey differential system:

$$\dot{q}_j(t) = \delta_{ji}\, r_{io}(t) + \sum_{\substack{l=1\\ l\neq j}}^{n} \frac{R_{jl}}{Q_l}\, q_l(t) - \left[\sum_{\substack{l=1\\ l\neq j}}^{n} \frac{R_{lj}}{Q_j} + \frac{R_{oj}}{Q_j} \right] q_j(t)$$

or, by integrating:

$$\int_0^\infty \dot{q}_j(t)dt = \delta_{ji} \int_0^\infty r_{io}(t)dt + \sum_{\substack{l=1\\ l\neq j}}^{n} \frac{R_{jl}}{Q_l} \int_0^\infty q_l(t)dt - \left[\sum_{\substack{l=1\\ l\neq j}}^{n} \frac{R_{lj}}{Q_j} + \frac{R_{oj}}{Q_j} \right] \int_0^\infty q_j(t)dt$$

Hence

$$q_j(\infty) - q_j(o) = \delta_{ji}\, q_{io} + \sum_{\substack{l=1\\ l\neq j}}^{n} \frac{R_{jl}}{Q_l} \int_0^\infty q_l(t)dt - \left[\sum_{\substack{l=1\\ l\neq j}}^{n} \frac{R_{lj}}{Q_j} + \frac{R_{oj}}{Q_j} \right] \int_0^\infty q_j(t)dt$$

The first member is nil, and consequently:

$$\delta_{ji} = - \sum_{l=1}^{n} k_{jl} \int_0^\infty \frac{q_l(t)}{q_{io}}\, dt \qquad \text{with} \qquad \delta_{ji} = \begin{cases} 0 \text{ if } j \neq i \\ 1 \text{ if } j = i \end{cases}$$

and

$$\int_0^\infty r_{io}(t)dt = q_{io}$$

The above set of relations for $j = 1, 2 \ldots n$ and $i = 1, 2 \ldots n$ is equivalent to the matrix relation

$$I = K \left\| \int_0^\infty \frac{q_j(t)}{q_{io}}\, dt \right\| \qquad \text{or} \qquad \left\| \int_0^\infty \frac{q_j(t)}{q_{io}}\, dt \right\| = K^{-1} = \Theta.$$

Finally, considering the basic relation (1d) and the two expressions (6a) and (7b) of the occupancy matrix, we obtain the two following and equivalent general expressions:

$$Q = \left\| \int_0^\infty \frac{q_j(t)}{q_{io}} \right\| R \tag{11a}$$

$$Q = \left\| \int_0^\infty \frac{q_j(t)}{q_{io}}\, dt \right\|^{-1} \left\| \int_0^\infty t\, \frac{q_j(t)}{q_{io}}\, dt \right\| R \qquad (11b).$$

The above matrix relations call for a first observation. They remain invariant if any one of the compartments is replaced by an arbitrary portion of itself. The relation (11a) may in fact be written as follows:

$$Q_j = \sum_{i=1}^{n} \int_0^\infty \frac{q_j(t)}{q_{io}}\, dt \cdot R_{io} \qquad j = 1, 2 \ldots n \qquad (11c)$$

Now, this relation does not change if Q_j is replaced by αQ_j (α included between 0 and 1), since as we are concerned with a perfect tracer $q_j(t)$ becomes $\alpha q_j(t)$. Thus, relations (11a) and (11b) are valid for an arbitrary portion S' of system S.

Further they have been well known and used for a considerable time, in various forms and in the particular case of a system with only one input, or only one input and output.

a. In the first case, the vector R represents a single component which we shall designate by R_o and relations (11c) may then be written:

$$Q_j = \int_0^\infty \frac{q_j(t)}{q_o}\, dt \cdot R_o$$

i.e. with consideration of any portion $Q_{S'}$ of the system and with reference to our earlier observation.

$$Q_{S'} = \int_0^\infty \frac{q_{S'}(t)}{q_o}\, dt \cdot R_o \qquad (12)$$

If this result is applied to a circulation system, the arbitrary pool $Q_{S'}$ is replaced by an arbitrary volume $V_{S'}$. If we introduce the radioactive concentration $C_{S'}(t) = \dfrac{q_{S'}(t)}{V_{S'}}$ the foregoing relation may be expressed as:

$$R_o = \frac{q_o}{\int_0^\infty C_{S'}(t)dt}$$

This is the old and famous relation of Stewart and Hamilton. $C_{S'}(t)$ is the concentration considered in an arbitrary region of the arterial system, radioactivity q_o being injected at any point of the veinous system, so that as all the radioactivity goes through the heart, the system may in fact be regarded as having a single input. R_o is then the flow through the heart.

If measurement of the above flow is by precordial detection of radioactivity, the relation is employed in the form:

$$R_o = \frac{q_o}{\int_0^\infty \frac{q_{S'}(t)}{V_{S'}}\, dt}.$$

If a is the counting geometry factor, measurement of precordial radioactivity gives us $aq_{S'}(t)$ and calibration gives us $aV_{S'}$. We therefore obtain the ratio $\frac{q_{S'}(t)}{V_{S'}}$. $V_{S'}$ is therefore the volume of blood "seen" by a suitably collimated detector. If we now consider the total pool Q_S of the organism for one element and $q_S(t)$ total radioactivity after injection of a suitable radioactive tracer, relation (11a) gives us:

$$Q_S = \int_0^\infty \frac{q_S(t)}{q_o} dt \cdot R_o$$

R_o is the daily intake of the element considered and is the mean residence time of the element, θ_S. This relation is also well known and makes it possible to determine the total pool of an element by whole body counting, if the radioactive tracer employed has a suitable half-life.

Relation (12), valid for an arbitrary portion S' of the system S and which may also be expressed:

$$Q_{S'} = \theta_{S'} R_o$$

is much less usual. It was formulated for the first time by Bergner (4) who was clearly aware of its general importance. In fact, it permits simple determination of the pool of a compartment, as understood by Bergner, who suggested its use for measurement of exchangeable mass (24). Its value for the solution of numerous problems involved in the kinetics of tracers has recently being stressed by Orr and Gillespie (25) who described what they call the occupancy principle. These researchers expressed the relation in the form:

$$\frac{1}{R} = \frac{\theta_{S'}}{Q_{S'}} .$$

They call $\theta_{S'}$ the occupancy and $Q_{S'}$ the capacity of the arbitrary compartment S'. The occupancy principle is then stated by saying that the ratio of occupancy to capacity is constant at any point of the system and equal to the inverse of the flow of matter R_o through the system. The value of this principle is essentially due to the fact that if the capacity of an arbitrary compartment S'_1 (such as a given volume of plasma) is known, determination of an occupancy states the daily intake. It is then sufficient to determine occupancy (by external detection, for example) of any other arbitrary compartment S'_2 of the organism to obtain the capacity or pool of this compartment S_2.

b. In the second case (a system S with a single input, and a single output), it is well known that:

$$Q = R_o t_S$$

with

$$t_S = \frac{\int_0^\infty t a(t) dt}{\int_0^\infty a(t) dt}$$

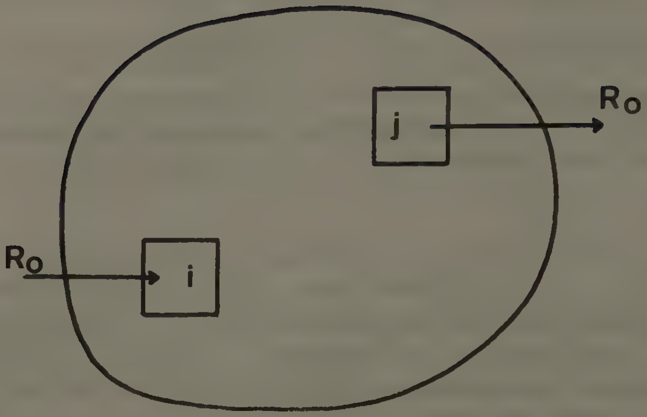


Fig. 2

whereby Q is the total pool of the system, R_o the flow of matter crossing it, t_S the mean transit time and $a(t)$ the specific radioactivity of the tracer at the system output. Let us assume that the input is at compartment i and output at compartment j (Fig. 2). q_{io} being the injected radioactivity at i and $q_j(t)$ the radioactivity at j at instant t, we may write:

$$t_S = \frac{\int_0^\infty ta(t)dt}{\int_0^\infty a(t)dt} = \frac{\int_0^\infty tq_j(t)dt}{\int_0^\infty q_j(t)dt} = \frac{\int_0^\infty t\,\frac{q_j(t)}{q_{io}}\,dt}{\int_0^\infty \frac{q_j(t)}{q_{io}}\,dt} = \left[\int_0^\infty \frac{q_j(t)}{q_{io}}\,dt\right]^{-1} \int_0^\infty t\,\frac{q_j(t)}{q_{io}}\,dt$$

Finally, the well-known relation

$$Q = \frac{\int_0^\infty ta(t)dt}{\int_0^\infty a(t)dt} R \tag{13}$$

which generally appears in the non-compartmental theory of tracer kinetics may be expressed as:

$$Q = \left[\int_0^\infty \frac{q_j(t)}{q_{io}}\,dt\right]^{-1} \int_0^\infty \frac{t\,q_j(t)}{q_{io}}\,dt \cdot R$$

In this form it is evidently a particular case of the general matrix relation (11b).

The general case.

Let us return to the general case of several inputs R_{io} and several distinct outputs R_{oi}. We have:

$$Q_j = \sum_{i=1}^{n} \theta_{ji} R_{io} \qquad \text{with } \theta_{ji} = \int_0^\infty \frac{q_j(t)}{q_{io}}\,dt.$$

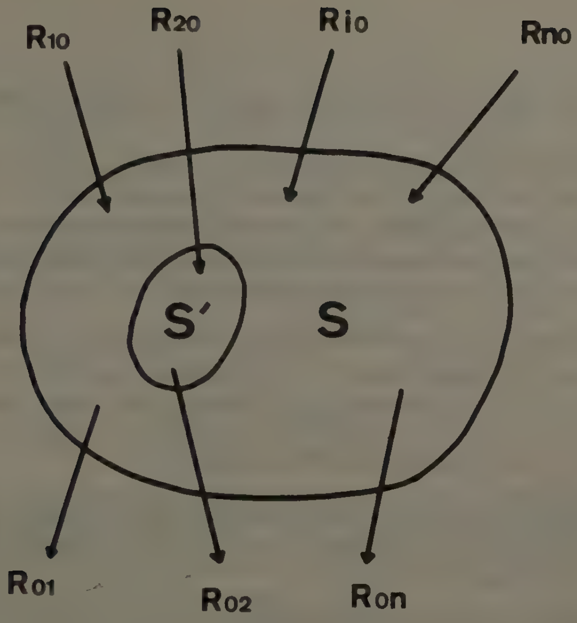


Fig. 3

We now consider a compartment S' of system S (Fig. 3). Its total pool may always be represented by the sum of Q_j or fraction αQ_j, j varying for example from l to m, with $1 \leqslant l < m \leqslant n$, the total pool is given by:

$$Q_{S'} = \sum_{j=l}^{m} \sum_{i=1}^{n} \theta_{ji} R_{io} = \sum_{i=1}^{n} R_{io} \sum_{j=l}^{m} \theta_{ji} = \sum_{i=1}^{n} R_{io}\, \theta_{S'i} = \sum_{i=1}^{n} R_{io} \int_0^\infty \frac{q_{S'}(t)}{q_{io}}\, dt$$

which we may formulate as follows:

$$Q_{S'} = R_o \sum_{i=1}^{n} \frac{R_{io}}{R_o}\, \theta_{S'i} = R_o \sum_{i=1}^{n} \frac{R_{io}}{R_o} \int_0^\infty \frac{q_{S'}(t)}{q_{io}}\, dt \tag{14}$$

$$R_o = \sum_{i=o}^{n} R_{io}$$

representing the total flow of matter entering the system. Two cases may then be considered:

a. if the total quantity of tracer injected q_o is redistributed in the various compartments of S in the same way as the flow of matter, we then have:

$$q_{io} = q_o \frac{R_{io}}{R_o}$$

and the foregoing relation may be written:

$$Q_{S'} = R_o \int_0^\infty \frac{q_{S'}(t)}{q_o}\, dt = \theta_{S'} R_o.$$

It is then said that there is *equivalent tracer supply*. Thus, the occupancy principle is presented in the form:

$$\frac{1}{R_o} = \frac{\theta_{S'}}{Q_{S'}}$$

for only one input extended to any number of inputs if there is equivalent tracer supply.

The most usual case of equivalent supply is that in which blood flow is measured by the Stewart and Hamilton principle. Due to the fact that whatever the point of tracer injection in the veinous compartment, the total volume of tracer must pass through a single point of the ramified circulation system, namely the heart, means that there is equivalent supply.

b. If there is no equivalent supply the relation (14) shows that to determine the pool $Q_{S'}$ of compartment S' it is necessary not only to determine each $\theta_{S'i}$ $(i = 1, 2 \ldots n)$ by injecting tracer successively at each input, but also to know the values $\frac{R_{io}}{R_o}$ of input flow distribution R_O.

Fortunately, the case of *several distinct inputs* remains a theoretical case, as in practice and for virtually all metabolism the various inputs, when there are several, lead into the plasma.

The same results are evidently applicable in the particular case of $S' = S$. We then have

$$Q_S = R_o \sum_{i=1}^{n} \frac{R_{io}}{R_o} \theta_{Si}.$$

Yet, according to relation (8b) we may then replace the mean time of sojourn or mean residence time θ_{Si} with the mean transit time t_{Si}. We may then write (8a), *in the case of impulse injection*

$$t_{ji} = \int_0^\infty t\, k_{oj}\, p_{ji}(t)\, dt = \int_0^\infty t\, \frac{R_{oj}}{Q_j} \frac{q_j(t)}{q_{io}}\, dt$$

if we call q_{oj} the total quantity of tracer eliminated from the system through compartment j and $a_j(t)$ the specific activity in the compartment, we have:

$$\int_0^\infty \frac{dq_{oj}}{R_{oj}} = \int_0^\infty a_j(t)dt = \frac{q_{oj}}{R_{oj}}$$

transit time t_{ji} may then be expressed as:

$$t_{ji} = \frac{q_{oj}}{q_{io}} \frac{\int_0^\infty t\, a_j(t)dt}{\int_0^\infty a_j(t)dt}$$

and we then have:

$$t_{Si} = \sum_{j=1}^{n} \frac{q_{oj}}{q_{io}} \frac{\int_0^\infty t\, a_j(t)dt}{\int_0^\infty a_j(t)dt}$$

and

$$Q_S = R_o \sum_{i=1}^{n} \frac{R_{io}}{R_o} t_{Si} = R_o \sum_{i=1}^{n} \frac{R_{io}}{R_o} \sum_{j=1}^{n} \frac{q_{oj}}{q_{io}} \frac{\int_0^\infty t\, a_j(t)dt}{\int_0^\infty a_j(t)dt} \qquad (15a).$$

This is the extension to n inputs and outputs of the classical relation (13) which is valid for only one input and one output. As can be seen, it is much more complex and less simple to use than the relation derived from (14)

$$Q_S = R_o \sum_{i=1}^{n} \frac{R_{io}}{R_o} \theta_{Si}$$

$$Q_S = R_o \sum_{i=1}^{n} \frac{R_{io}}{R_o} \int_0^\infty \frac{q_S(t)}{q_{io}} dt \qquad (15b).$$

This is due to the fact that the form (11b) of the matrix Θ is less simple than the form (11a). Nonetheless, the relation (15a) permits determination of the total pool of the organism (system S) by measurements made on samples, whereas relation (15b) demands whole-body counting.

In practice, things become simpler. As a rule, the input is single (the plasma) and the relation (15a) is written as:

$$Q_S = R_o \sum_{j=1}^{n} \frac{q_{oj}}{q_o} \frac{\int_0^\infty t\, a_j(t)dt}{\int_0^\infty a_j(t)dt} \qquad (15c)$$

Further, the number of values of j is frequently reduced to 2 (urine and feces), and will be no higher than 4 if pulmonary and cutaneous elimination are considered.

Turnover time and turnover rate.

We saw in the section concerning representation of a system by a Markov process that such magnitudes as occupancy θ_{ji} or transit time t_{ji} have the dimensions of a time, but do not represent a mean value of this variable. It does not therefore seem that these times can correspond to a turnover time and their inverse to a turnover rate.

Contrarily, time θ_{ii} corresponding to injection of tracer in compartment i itself represents a mean value. This is the *mean time of sojourn, mean residence time* or *turnover time* of particles in compartment i. His inverse represents the *turnover rate* of matter in this compartment.

Generally speaking in the case of an arbitrary compartment S', corresponding to a number of compartments or part of compartments Q_j, we obtain the following result:

If tracer injection is given in a compartment i of system S, external to S', the occupancy $\theta_{S'i}$ do not represent a mean time and do not correspond with any turnover time or turnover rate.

If contrarily injection is given in a compartment i belonging to S', the occupancy $\theta_{S'i}$ represents a mean time. It is the mean time of sojourn, mean residence time or turnover time of particle in compartment S', which are entering into the system by input R_{io}. His inverse represents the turnover rate of these particles in the compartment S'.

In the case of the whole system S

$$\theta_{Si} = t_{Si}$$

t_{Si} represents a mean time, i.e. the mean transit time through the system of the particles entering at input R_{io}. Its inverse corresponds to the turnover rate of all the particles in the system which entered at R_{io}, since $t_{Si} = \theta_{Si}$.

In the final analysis

$$\theta = \sum_{i=1}^{n} \frac{R_{io}}{R} \theta_{Si} = \frac{R_{io}}{R} t_{Si} = t_S$$

represents the turnover time of the whole system and its inverse the system's turnover rate.

These results are schematically represented in figure 4.

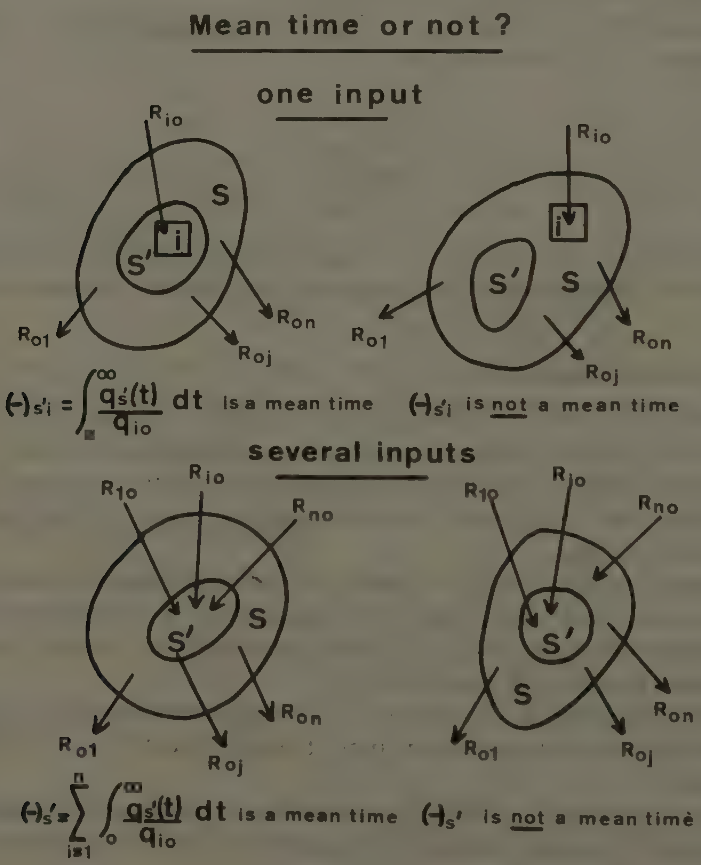

Mean time or not ?
one input
R_{io}
S
S'
i
R_{on}
R_{o1}
R_{oj}
$\Theta_{S'i} = \int_0^\infty \frac{q_{S'}(t)}{q_{io}} dt$ is a mean time
R_{io}
i
S'
S
R_{o1}
R_{on}
R_{oj}
$\Theta_{S'i}$ is not a mean time
several inputs
R_{1o}
R_{io}
R_{no}
S
S'
R_{on}
R_{o1}
R_{oj}
$\Theta_{S'} = \sum_{i=1}^{n} \int_0^\infty \frac{q_{S'}(t)}{q_{io}} dt$ is a mean time
R_{1o}
R_{io}
R_{no}
S'
S
R_{on}
R_{o1}
R_{oj}
$\Theta_{S'}$ is not a mean time

Fig. 4

TYPICAL APPLICATIONS

The following is a brief description of a few examples bearing essentially on applications of the occupancy principle. They are mostly derived from the work done jointly by the Department of Clinical Physics and Bioengineering of the Western Regional Hospital Board, Glasgow, and the Service Hospitalier Frédéric Joliot of the Commissariat à l'Energie Atomique, at Orsay (26).

Measurement of intra-thyroid iodine.

Figure 5 shows the thyroid radioactivity curve of a normal subject as measured by external detection, and the radioactivity curve of the organic iodine compartment per liter of plasma, after ingestion of 325 μg of ^{129}I representing 0.05 μCi. The two curves are expressed in percentage of ingested radioactivity and cover a period of nearly 240 days.

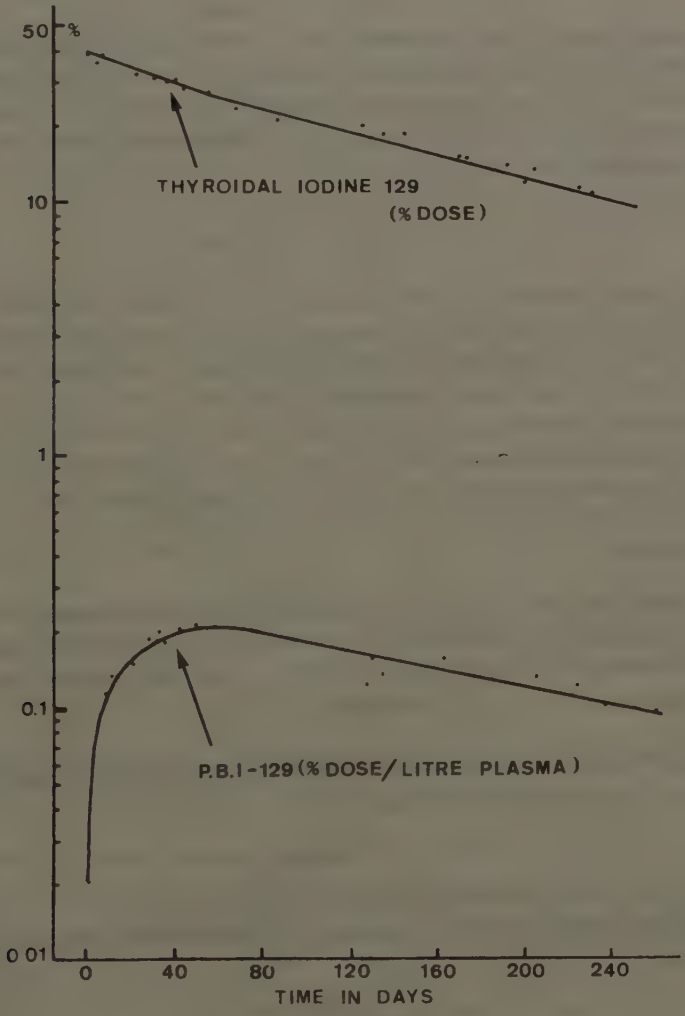


Fig. 5

They show an occupany θ_T of 68.1 days for the thyroid, and an θ_P of 0.52 days for PBI in one liter of plasma. Further, the quantity of organic iodine per liter of plasma Q_P measured by activation analysis is 43 μg.

R_o being the daily intake of iodine, the occupancy principle establishes that:

$$\frac{1}{R_o} = \frac{\theta_T}{Q_T} = \frac{\theta_P}{Q_P}$$

whereby we deduce the quantity of iodine in the gland

$$Q_T = Q_P \frac{\theta_T}{\theta_P} = \frac{43 \cdot 68.1}{0.52} = 5{,}600 \ \mu g$$

and the daily intake:

$$R_o = \frac{Q_P}{\theta_P} = \frac{43}{0.52} = 82 \ \mu g/\text{day}.$$

It will be seen that the plasma radioactivity curve goes through a maximum representing 0.0021% of the ingested radioactivity, per liter. At the same time, thyroid radioactivity is 0.27% of ingested radioactivity. If it is then accepted that the specific radioactivities in the two successive compartments, thyroid and plasma, are the same, we deduce that:

$$Q_T = 0.27 \cdot \frac{43}{0.0021} = 5{,}500 \ \mu g.$$

The two values obtained for the thyroid pool by these two quite different methods are in this case in excellent agreement. This is by no means the case for endemic goiters, in which the second method produces values that may be two or three times smaller than those produced by the first. This is due to the fact that unlike euthyroid subjects, goiters involve a thyroid gland which cannot be considered as a single compartment. In the latter case, measurement of the thyroid iodine pool by the occupancy principle is the only valid measurement.

Rubidium metabolism.

Figure 6 shows the total radioactivity curves expressed in percent of ingested radioactivity and relating to a human subject, as measured by whole body counting, and those showing the radioactivity per liter of saliva measured with a well crystal, after ingestion of 70 μCi of Rubidium 86.

Measurement of the daily intake of rubidium by activation analysis for a period of 40 days provided a mean value :

$$R_o = 4.32 \ \text{mg/day}$$

and shows that the subject was virtually in the steady state.

The foregoing curves give us for θ_B and θ_S of whole body occupancy and in one liter of saliva :

$$\theta_B = 43 \ \text{days} \qquad \theta_S = 0.35 \ \text{days}.$$

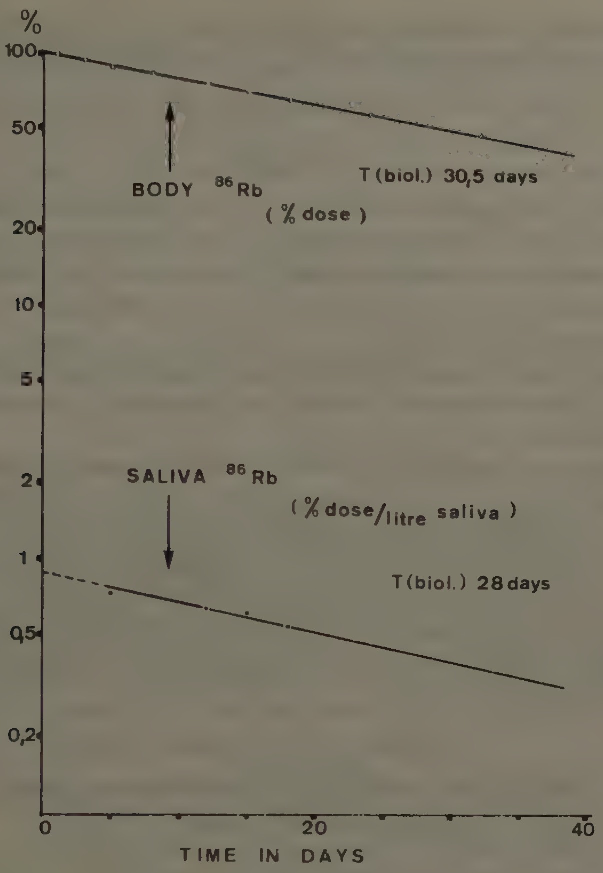


Fig. 6

Moreover, the quantity of rubidium per liter of saliva measured by activation analysis is $Q_S = 1.51$ mg/liter.

The occupancy principle: $\frac{1}{R_o} = \frac{\theta_S}{Q_S} = \frac{\theta_B}{Q_B}$ give $Q_B = R_o\theta_B = 4.32 \cdot 43 = 185.7$ mg

and $Q_B = \theta_B \frac{Q_S}{\theta_S} = 43 \cdot \frac{1.51}{0.35} = 185.5$ mg.

The two values are in excellent agreement and confirm the validity of the occupancy principle.

Change from one steady-state to another.

Let R_{ji} be the transport into compartment j from i, Q_i the pool of compartment i, and k_{ji} the corresponding transport rate. The system being in the steady state, we have

$$R_{ji} = k_{ji}\, Q_i.$$

If this transport mechanism may be considered as corresponding to a reaction of order *n* with reaction constant k'_{ji}, we obtain by definition :

$$R_{ji} = k'_{ji}\, Q_i^n.$$

Hence (2)

$$k_{ji} = k'_{ji}\, Q_i^{n-1}.$$

We see that if the transport mechanism is a reaction of an order > 1, any disturbance of the steady state producing a change in the value of pool Q_i, will involve a concommitant variation k_{ji}, since k'_{ji} is constant by definition. In such case, any perturbation of the steady state will involve a modification of the element θ_{ji} of the matrix Θ.

Contrarily, if the transport mechanism corresponds to a reaction of the first order $k_{ji} = k'_{ji}$ remains constant whatever the changes in Q_i during perturbation. The same applies to θ_{ji}.

The occupancy principle thus offers the by no means negligible advantage of independence with respect to fluctuations of the system's steady state within the limits of first order kinetics. Measurements employing "occupancy" produce results which are relatively independent of biological fluctuations of the steady state, insofar as these generally follow kinetics of the first order.

a. Figure 7 show two schematics of three compartments each and representing two different steady states of iodine metabolism in the case of a markedly hyperthyroidian subject.

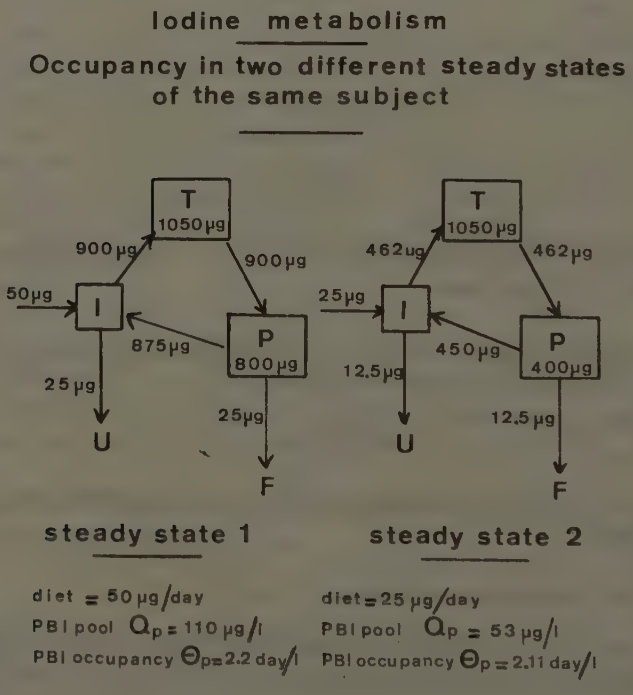

Iodine metabolism
Occupancy in two different steady states of the same subject
T
1050 µg
900 µg
900 µg
50 µg
I
875 µg
P
800 µg
25 µg
25 µg
U
F
T
1050 µg
462 µg
462 µg
25 µg
I
450 µg
P
400 µg
12.5 µg
12.5 µg
U
F
steady state 1
steady state 2
diet = 50 µg/day
PBI pool Qp = 110 µg/l
PBI occupancy Θp = 2.2 day/l
diet = 25 µg/day
PBI pool Qp = 53 µg/l
PBI occupancy Θp = 2.11 day/l

Fig. 7

The first relates to a daily iodine intake of 50 μg per day. The intake was abruptly reduced from 50 to 25 μg per day. The steady state was thus terminated and the subject developed towards a further steady state, reached within 30 days.

The second steady state is shown in the second schematic. In the first steady state, plasma iodine concentration was 110 μg per liter, and the occupancy per liter of plasma was therefore :

$$\theta_T = \frac{110}{50} = 2.20 \text{ days.}$$

In the second steady state, reached after changing the intake from 50 to 25 μg per day, plasma concentration became 53 μg per liter and occupancy:

$$\theta_P = \frac{53}{25} = 2.11.$$

It will be noted that plasma occupancy seems to have changed very little during transition from one steady state to the next. This leads to supposition that the transport mechanisms of the subject in the plasma must follow kinetics of the first order. This probably applies in particular to peripheral de-halogenization. Contrarily, occupancy is in this case substantially changed in the thyroid gland, whose pool appears to remain practically constant during the transitional period. These assumptions afforded correct interpretation of the variation of specific radioactivity during the change of state.

b. An experiment was made with a group of 20 rats, each of which was brought to the steady state with a different intake. The intakes ranged between 3 and 150 μg (per day). For each rat, both occupancy of the thyroid and that of the thyroxine were measured. The occupancy principle permits determination of the intra-thyroid iodine pool of each animal.

Figure 8 shows the variation of thyroid and thyroxine occupancy, and the quantity of iodine in the thyroid as a function of daily intake. It can be seen that at a more than 10 μg per day (which is a considerable load of iodine for a rat) and up to 150 μg per day, the intrathyroid iodine pool remains perfectly constant. The gland seems saturated. Occupancy of the thyroid and the thyroxine diminish in parallel as the intake is increased, the product of the two magnitudes (thyroid iodine pool) remaining constant.

It may thus be considered that between 10 and 150 μg per day of iodine, the magnitude of the overload for a rat triggers active physiological regulatory mechanisms which are very different from first order kinetics.

Regarding intakes of less than 10 μg per day, there are only two experimental data showing that the intrathyroid iodine pool decreases rapidly as the intake is reduced. Exact experiments in this field are being conducted to see whether or not it is the thyroid capacity which diminishes with intake, leaving occupancy relatively unchanged. Such a result would be fairly logical, if we may expect that small excesses of iodine cause perturbations of the first order of kinetics.

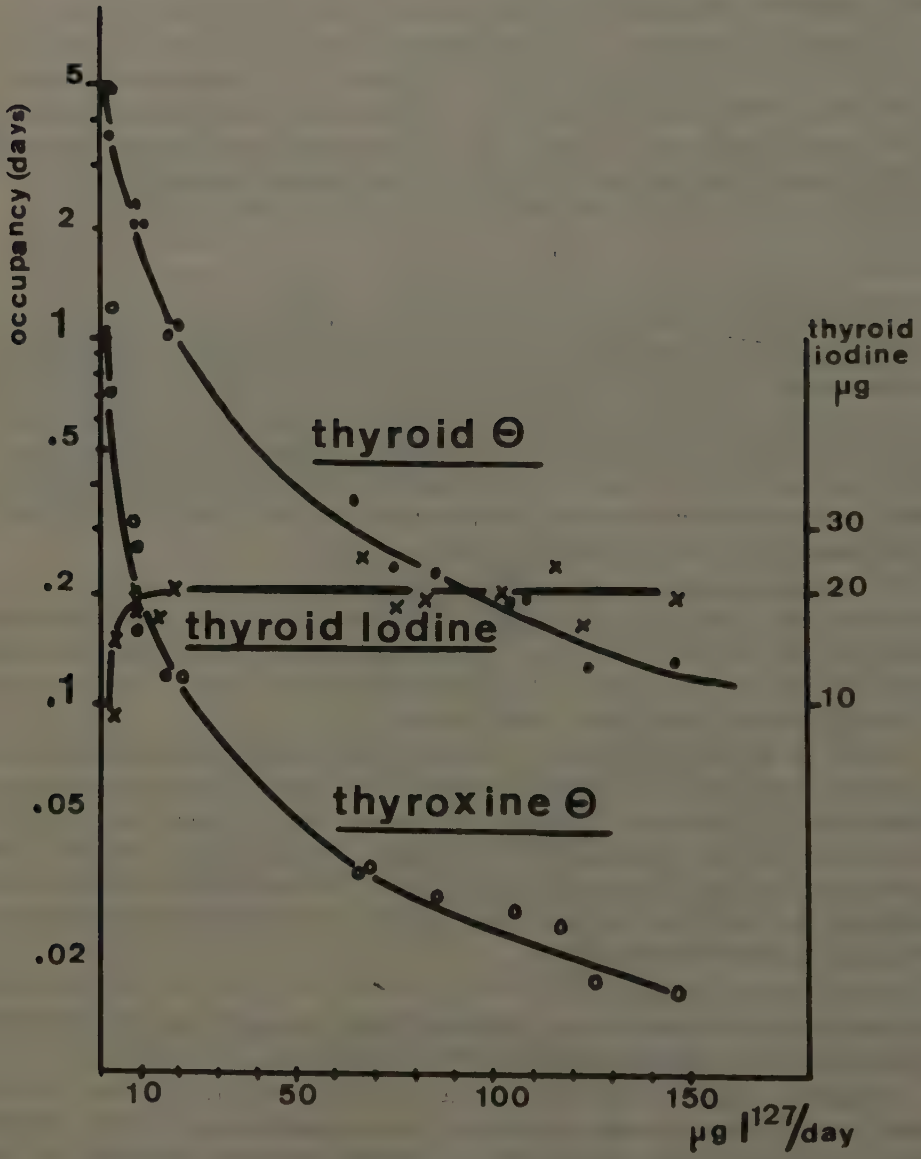

occupancy (days)
5
2
1
.5
.2
.1
.05
.02
thyroid Θ
thyroid iodine
thyroxine Θ
thyroid iodine µg
30
20
10
10
50
100
150
µg I127/day

Fig. 8

REFERENCES

1. Von Hevesy, G., "Radioactive indicators, their applications in biochemistry, animal physiology and pathology", Interscience, New York 1948.
2. Bergner, P. E. E., "Tracer dynamics: I. A tentative approach and definition of fundamental concepts", *J. Theoret. Biol.* **2,** 120, 1961.
3. Bergner, P. E. E., "Tracer dynamics: II. The limiting properties of the tracer system", *J. Theoret, Biol.* **1,** 359, 1961.
4. Bergner, P. E. E., "Tracer dynamics and the determination of pool-sizes and turnover factors in metabolic systems", *J. Theoret. Biol.* **6,** 137, 1964.
5. Solomon, A. K., "The kinetics of biological processes: special problems connected with the use of tracers", *Adv. Biol. Med. Phys.* **3,** 65, 1953.
6. Robertson, J. S., "Theory and use of tracers in determining transfer rates in biological systems", *Physiol. Rev.* **37,** 133, 1957.
7. Solomon, A. K., "Compartmental methods of kinetic analysis", In "Mineral Metabolism" (Ed. Comar, C. L. and Bronner, F.) Vol. A, Pt A., pp. 119-67, Academic Press, New York, 1960.
8. Zierler, K. L., "Circulation times and the theory of indicator dilution methods for determining blood flow and volume", Handbook of physiology, Sec. 2, Vol. 1, pp. 586-615, Am. Physiol. Soc., Washington, D.C., 1962.
9. Robertson, J. S., "Mathematical treatment of uptake tracers and release of indicator substances in relation to flow analysis in tissues and organs", Handbook of physiology, Sec. 2, Vol. 1, pp. 617-644, Am. Physiol. Soc. Washington, D.C., 1962.
10. Berman, M., "Compartmental analysis in kinetics", In "Computers in Biomedical Research", Vol. 2, Chapter VII (Ed. Stacy, R. and Waseman, B.) Academic Press, New York, 1965.
11. Bergner, P. E. E., "Tracer theory: a review", *Isotop. Radiat. Technol.* **3,** 245, 1966.
12. Brownell, G. L., Berman, M. and Robertson, J. S., "Nomenclature for tracer kinetics" *Int. J. appl. Rad. Isotopes* **19,** 249, 1968.
13. Manougian, E., "Kinetics of metabolic processes as studied with radioactive tracers", Progress in Atomic Medicine, **2,** 184, 1968.
14. Sheppard, C. W., "Basic principles of the tracer method", Wiley, New York, 1962.
15. Rescigno, A. and Segre G., "Drug and tracer kinetics", Blaisdell, Waltham, Mass., 1966.
16. Hart, H. E., (Ed.) "Multi-compartment analysis of tracer kinetics", Ann. N.Y. Acad. Sci., 1963.
17. Kniseley, R. M. and Tauxe, W. N., (Eds.) "Dynamic clinical studies with radioisotopes", AEC-3 Symposium series, U.S. Atomic Energy Commission, Div. of Technical Information, TID 7678, 1964.
18. Bergner P. E. E. and Lushbaugh C. C., (Ed.) "Compartments, pools and spaces in medical physiology". AEC Symposium Series, U.S. Atomic Energy Commission, Div. of Technical Information, CONF-661010 1967.
19. Berman, M. and Schoenfeld, R., "Invariants experimental data on linear kinetics and the formulation of models", *J. appl. Phys.* **27,** 1361, 1956.
20. Cox, D. R. and Miller, H. D., "The theory of stochastic processes", Methuen London, 1965.
21. Feller, W., "An introduction to probability theory and its applications", Vol. I, 3rd. Wiley, New York, 1968.
22. Hearon, J. S., "Theorems in linear systems" in ref. 16 pp. 36.
23. Hart, H. E., "Analysis of tracer experiments in non-conservative steady-state systems", *Bull. Math. Biophysics* **17,** 87, 1955.
24. Bergner, P. E. E., "Exchangeable mass: determination without assumption of isotopic equilibrium", *Science,* **150,** 1048, 1965.
25. Orr, J. S. and Gillespie, F. C., "Occupancy principle for radioactive tracers in steady-state biological systems", *Science* **162,** 138, 1968.
26. Riviere, R., Comar, D., Kellershohn, C., Orr, J. S., Gillespie, F. C. and Lenihan, J. M. A., "Estimation of thyroid iodine content by the occupancy principle", The Lancet, Feb. 22, 1969, 389.

MODULATION TRANSFER FUNCTION AND ITS APPLICATION TO RADIOISOTOPE SCANNING

T. D. Cradduck

Introduction

Since the first radioisotope scanners were developed by Cassen, et al. (1) and Mayneord, et al. (2) there has been a continuous interest in methods which could be used to adequately assess and characterize the physical parameters of the systems involved. The major parameter which has been given this attention is that of RESOLUTION or RESOLVING POWER, which has been given the name DETECTABILITY in a recent I. C. R. U. report (3). Another major parameter is that of SENSITIVITY, which is, of course, another factor affecting DETECTABILITY.

The concepts which have been applied to detectability have been varied but most of them have some derivation in the field of optics. It is not surprising, therefore, to find that the earlier methods of defining detectability are analogous to the earlier methods of defining the resolving power of optical instruments. The concept of defining the resolviing power of a telescope in terms of its ability to resolve doublets leads naturally to the definition of detectability for a scanner in terms of the separation of two point sources.

Due to our background of radiotherapy it was not surprising that isoresponse curves soon came to be the order of the day in describing collimator response to a point source. The analogy to isodose curves for therapy machines is immediately apparent. In order to provide a single number describing the detectability of the collimators it was a natural step to define this as the width of the 50% isoresponse curve or, alternatively the full width at half maximum (FWHM) of the curve obtained when a point source was scanned across in front of the collimator in the focal plane.

At the same time as these developments were taking place, so also, were developments being made in the fields of optics and diagnostic radiology. It was becoming increasingly evident that the methods of describing resolution or detectability were subject to criticism and new concepts were being introduced in these two areas. The most notable amongst these concepts was that of the modulation transfer function (4, 5, 6, 7) and Perrin (8) has written a review of the history of the introduction of this concept to the field of optics. The concept of modulation transfer functions has been applied to diagnostic radiology by Morgan (9, 10, 11, 12), Moseley (13), Rossmann (14, 15) and Doi (16, 17). Beck (18) was one of the first persons to apply the concept to radioisotope scanning and this subject was the source of considerable discussion at a Symposium held in Chicago in 1965 (19).

Modulation Transfer Function

The modulation transfer function is analogous to the frequency response of a hi-fi system or tape recorder. It describes the response of the scanning system to a spectrum of frequencies, though in the case of scanners these frequencies are spatial rather than temporal in nature. In more detail, if a scanning system scans an object, the activity within which is varying sinusoidally, the output of the system will also be sinusoidal but of different amplitude. The ratio of the output modulation to the input modulation is a measure of the modulation transfer coefficient at that particular frequency and a plot of the variation of this ratio against frequency constitutes the complete modulation transfer function (Fig. 1). Should there be a phase reversal between input and output this will exhibit itself as a negative modulation transfer coefficient. Such a phase reversal results in spurious resolution as described by Beck (18).

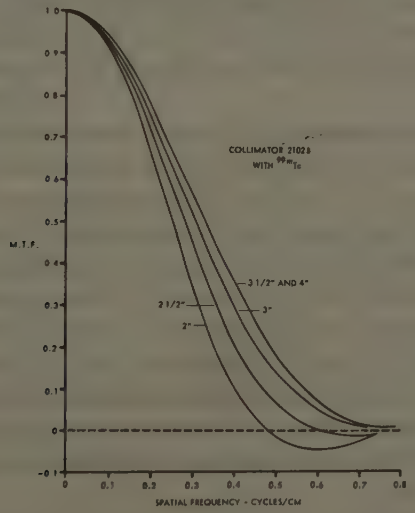


Fig. 1. Example of M.T.F. curves for several different source-to-collimator face distances. The curves for 2 and 2½ inch depths both demonstrate phase reversal which would result in spurious resolution for the frequencies involved (about 0.45 lines/cm and above).

One important factor is that the system be linear, i.e. that the ratio of output to input amplitudes remain constant for all (or, at least, a very large range) of the input amplitudes. This condition is met by most scanning data acquisition systems, though in most cases the data handling and presentation systems tend to be nonlinear.

The importance of linearity has been ably demonstrated by Mozley (19). He has attempted to measure the M.T.F. of a total scanning system including the data handling and presentation components. Due to the inherent nonlinearity of these two components (and particularly of the film) Mozley obtained an overall M.T.F. which was better than the M.T.F. of the data acquisition component. Although such a result may obtain it does so under only very limited circumstances dependent upon many factors such as the gamma curve of the film and consequently does not have general applicability.

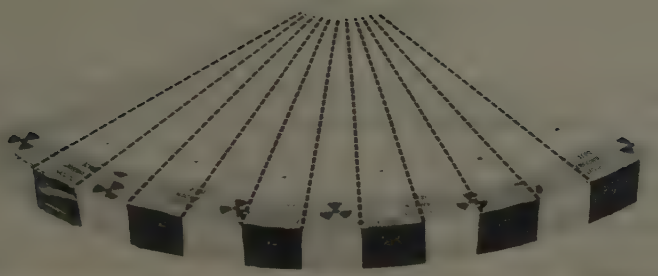
Similarly, the determination of the M.T.F. of image intensifier type stationary detector scanners presents some difficulty due to the restricted range of linearity of these systems. Recent technical advances have made the measurement of the line source response function from such devices possible so that we may expect to see some data relating to image intensifier systems in the near future and the range of linearity will become more clearly defined.

Measuring M.T.F.

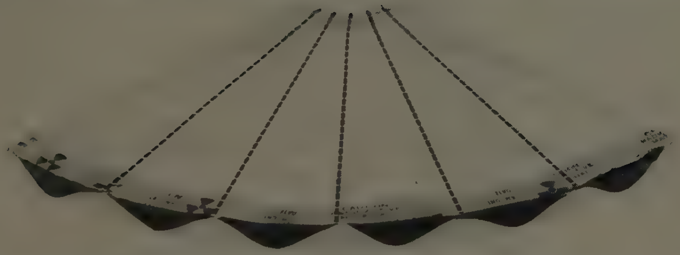
The first method described by Beck (18) for determining the M.T.F was to use an adaptation of the Siemen's star. A phantom was made up of a number of narrow wedges alternately containing activity or devoid of activity, (Fig. 2A). By recording the count rates over areas of activity and no activity at the same radius it was possible to measure the modulation transfer coefficient at the frequency corresponding to that radius. Repeated measurements at several radii resulted in a complete M.T.F.

The major disadvantage of this technique was that the phantom was of crenelate or square-wave form, so that the M.T.F. was somewhat enhanced since the phantom consisted not only of the fundamental frequency, but also the odd harmonics thereof. A modification of this technique was to use a sinusoidally varying wedge star phantom (Fig. 2B) which reduced the enhancement of the M.T.F. due to the square-wave phantom (21). However, an inherent disadvantage still existed in that the field of view of the collimator will extend over a range of radii, and, therefore, over a range of spatial frequencies, so that the measured modulation transfer coefficient at each setting is really an average figure and minor perturbations tend to be overlooked. Another disadvantage is that this type of phantom is difficult to manufacture, particularly if one chooses to use the rather more sophisticated sinusoidal phantom.

Fortunately, the mathematical analysis of the modulation transfer function leads to a simpler technique which has now been almost universally adopted (3, 21). Since the modulation transfer function or M.T.F. represents the system's response to all frequencies, if a source can be found which contains all frequencies, a Fourier frequency analysis of the output resulting from such a source will



A



B

Fig. 2A. The upper illustration shows a Siemen's star wedge phantom of the crenelate or square wave type.

Fig. 2B. The lower illustration depicts a Siemen's star phantom which has been made with a sinusoidal variation in depth.

give the M.T.F. A delta function or line source does contain all frequencies so that the Fourier analysis of a line source response function (LSRF) will give the M.T.F. (22).

Hine (23) and the I.C.R.U. Task Group (3) have suggested the use of a line source less than 1 mm. diameter and 30 cms. long suspended in a yoke 15 cms. radius as the most suitable for the determination of the line source response functions.

If $L(x)$ represents the line source response function then the M.T.F is given by:

$$MTF(\nu) = \frac{\int_{-\infty}^{+\infty} L(x) \cos 2\pi\nu x \, dx}{\int_{-\infty}^{+\infty} L(x) \, dx} \qquad \ldots 1$$

If the line source response function $L(x)$ is measured at a number of discrete points then the integration given in equation 1 can be expressed as a summation:

$$MTV(\nu) = \frac{\sum\limits_{i=n,m} L(x_i) \cos 2\pi\nu x_i}{\sum\limits_{i=n,m} L(x_i)} \qquad \ldots 2$$

and a computer can very readily be programmed to carry out the summation, multiplication and division implicit in this equation (21). Only the cosine term of the Fourier analysis is normally considered since the line source response function, $L(x)$ is usually assumed to be symmetrical about the $x = 0$ axis and the sine terms consequently drop out.

Value of the M.T.F.

That the M.T.F. is a valuable means of expressing the detectability of a scanning system can be seen if consideration is given to the line source response functions which are used to derive the M.T.F.'s. Figure 3 shows two line source response functions which have the same full width at half maximum. However, the

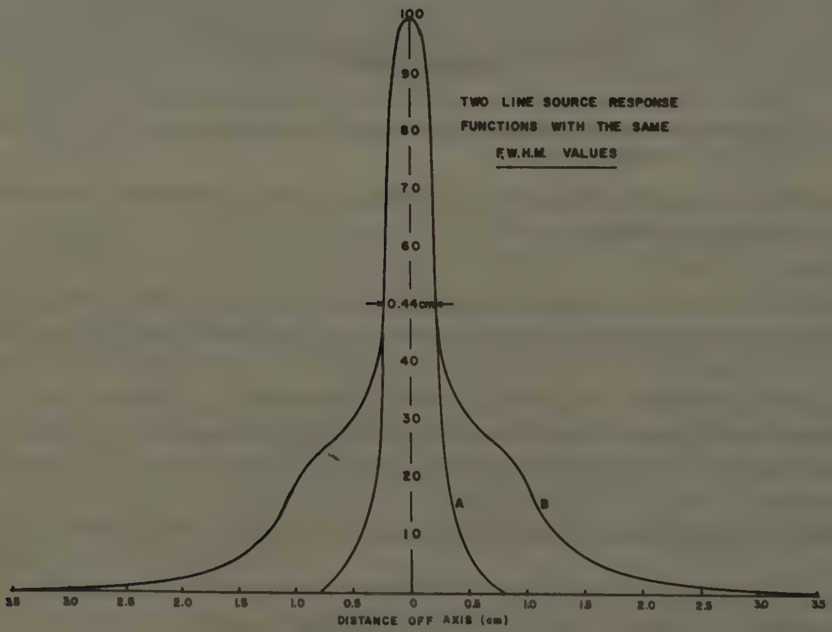


Fig. 3. Two hypothetical line source response functions having similar full widths at half maximum but exhibiting different characteristics below the 50% level.

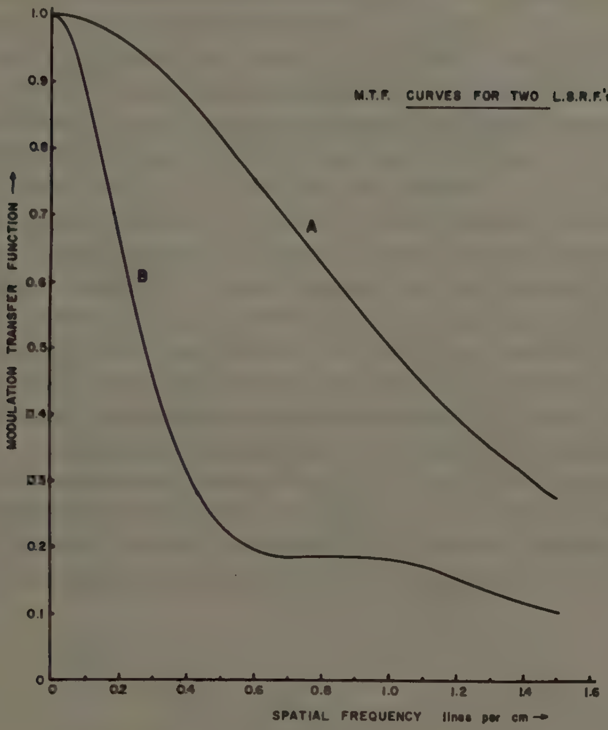


Fig. 4. Two M.T.F. curves resulting from the line source response functions shown in Figure 3. The large difference between the two curves results from the difference between the two line source response functions below the 50% level.

M.T.F.'s for these two line source response functions will be markedly different as shown in Fig. 4 due to the long tails of the second line source response function.

If only the full widths at half maximum were quoted for the two line source response functions depicted it would leave the impression that the two collimators would have similar characteristics. That this is not the case is very apparent from the M.T.F.'s for the two cases.

Another advantage of the use of the M.T.F. to characterize a scanning system is that the M.T.F. of each part of the system can be determined and then the separate M.T.F.'s can be multiplied together to give the overall M.T.F. As an instance of this Frey (24) and Gopala Rao et al. (25) have calculated the M.T.F. of the ratemeter system of a scanner based upon the ratemeter time constant.

The same methods have been used to determine the M.T.F. of the detecting crystal, photo-multiplier array and pulse forming network of a scintillation camera in the absence of any collimator (26, 27); this being equivalent to a measure of the inherent resolution of the detecting system. Keller et al. (28) have adopted the same concept when investigating the overall imaging capabilities in gamma-ray imaging devices.

A comprehensive library of modulation transfer functions is now being accumulated by a number of investigators working in collaboration with the International Atomic Energy Agency. Measurements have been made on a large range of the Picker-Nuclear collimators and some Selo collimators by Hine and Tsialas (29), on each of the Nuclear Chicago collimators and some of the Ohio-Nuclear five-inch collimators by Ross and Cradduck (30, 31), on the Ohio-Nuclear five- and eight-inch collimators by Christie and MacIntyre (32), and on the collimators made by three Japanese manufacturers by Iio (33). Without suggesting that the technique is by any means simple, it has arrived at the stage where a series of line source response functions can be used as input data for a computer program (34) and the output will be in the form of a series of line source response functions and M.T.F.'s all properly plotted out on the correct axes and suitably titled to comply with the suggestions of the I.A.E.A. group and its consultants (Fig. 5). Mallard (35) has suggested that the method be adopted to evaluate all collimators not just those used with scanners and the method is gaining wider acceptance.

Some investigators have noted that under some circumstances they obtain M.T.F. curves which exhibit minor perturbations resembling oscillations instead of the smooth curve normally expected. An example of such an M.T.F. is shown in Fig. 6. Sanderson (36) has suggested two possible sources of such perturbations.

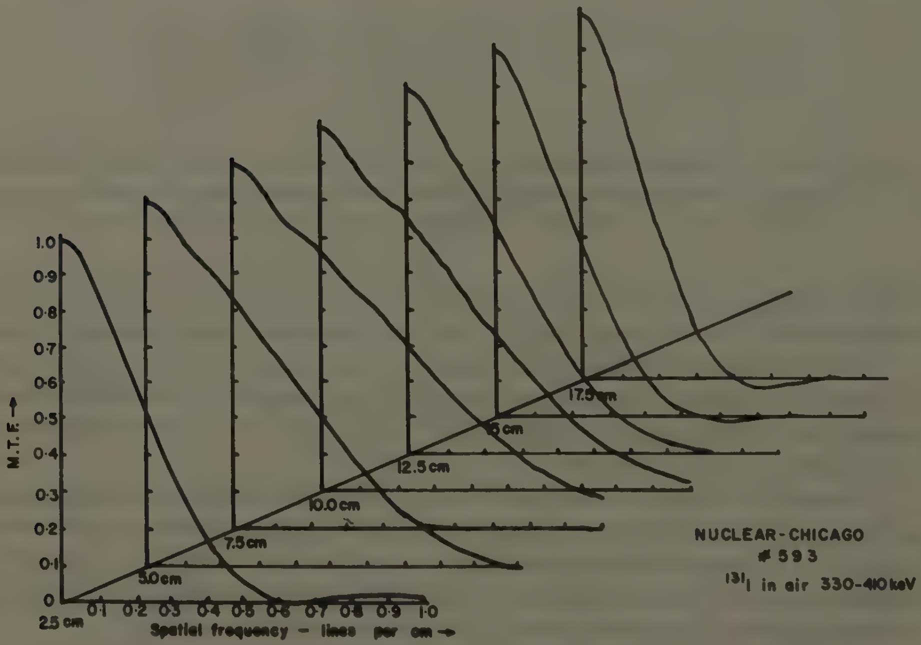


Fig. 5. A sample of the standard format adopted by the I.A.E.A. for plotting collimator M.T.F.'s. This illustration is drawn directly from the computer output of a program originating at the I.A.E.A. Computer Sciences Division. Some scale labelling has been omitted for the sake of clarity.

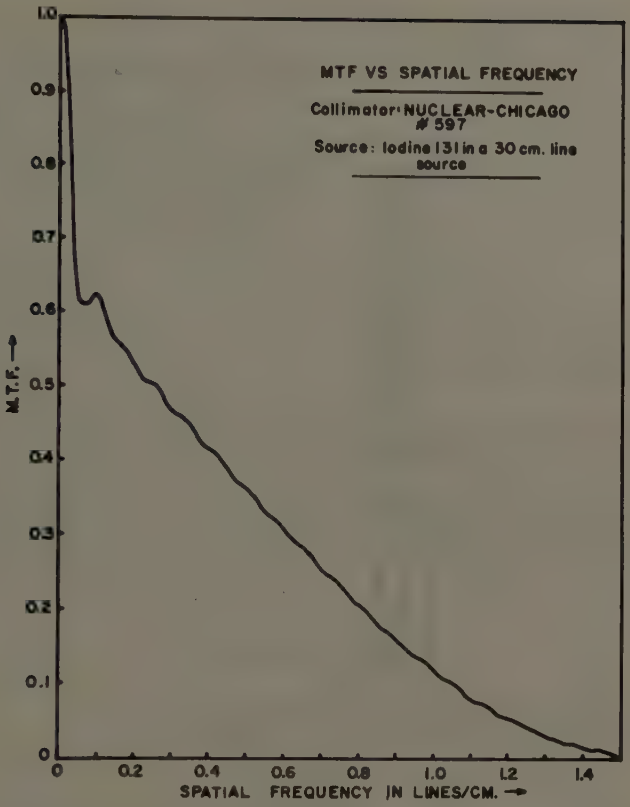


Fig. 6. The M.T.F. curve for the Nuclear-Chicago #597 collimator taken at the focal depth indicating the minor perturbations that can sometimes be expected.

The first is that the measurement of the line source response function is cut off too early and that consequently the M.T.F. calculation suffers from having finite instead of infinite limits applied to the integration. The second cause which Sanderson suggests is the presence of minor perturbations or irregularities in the line source response function itself. In the example cited in Fig. 6 the corresponding line source response function exhibits both of these characteristics. It can be seen (Fig. 7) that the line source response function close to the focal depth has minor irregularities close to the shoulder of the curve and, additionally, the measurement has not been carried out to the point at which the line source response function becomes zero.

Scatter and Penetration Fractions

Scatter: By postulating that the scattering and penetration processes can be considered as stages in the image forming system Beck et al. (37) have demonstrated the applicability of the modulation transfer function to measurements of the

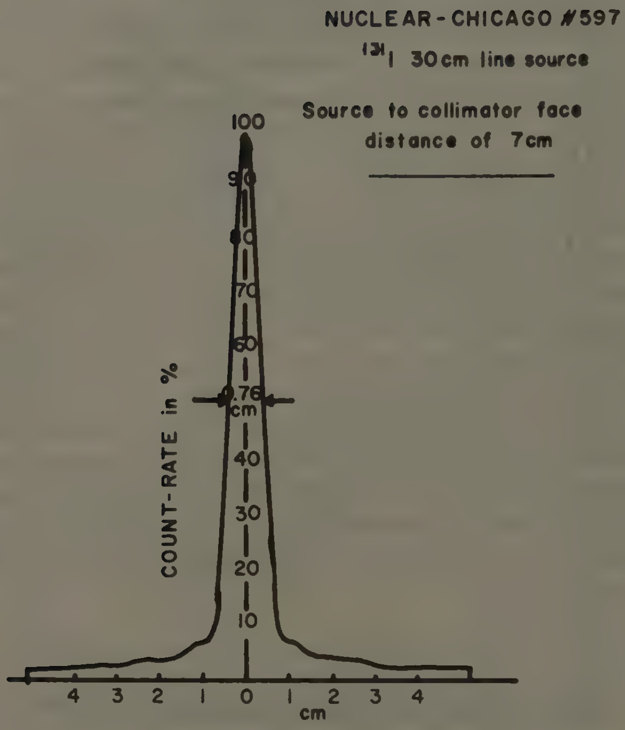


Fig. 7. The line source response function which was used to give the M.T.F. curve of Figure 6. A minor perturbation is observed at the shoulder of the curve and the sampling was cut off before the values reached zero. Both of these factors can give rise to perturbations in the M.T.F.

scatter fraction and this application could perhaps be further extended to include measurements of the penetration fraction. The scatter and penetration fractions are defined as the ratio of scattered (or penetrating) radiation to primary radiation reaching the crystal and being included as acceptable data.

Adopting Beck's nomenclature, the modulation transfer function may be expressed as:

$$MTF(\nu) = \frac{MTF_g(\nu) + P.MTF_p(\nu) + S.MTF_s(\nu)}{1 + P + S} \qquad \ldots 3$$

Where P and S are the penetration and scatter fractions respectively.

Let us, first of all, consider the situation when we have a collimator with relatively thick septa and a line source containing a lower energy radionuclide. Under these circumstances it may be reasonably assumed that penetration is negligible ($P = 0$) and equation 3 reduces to:

$$MTF(\nu) = \frac{MTF_g(\nu) + S.MTF_s(\nu)}{1 + S} \qquad \ldots 4$$

Let us further assume that the measurement of the line source response function is made in air so that scatter may also be neglected. Under these circumstances the line source response function will be almost Gaussian in shape as shown in Fig. 8.* As soon as a scattering medium is introduced the line source response function will be changed significantly and will have long tails due to the radiation scattered into the collimator from the source both within and outside of the collimator field of view (Fig. 8). The difference between these two line source response functions may be regarded as the line source response function due to scattered photons ($L_s(x)$) and this is also shown in Fig. 8.

At this point it is worthy of note that the FWHM values for each of the line source response functions are essentially the same despite the marked dissimilarity between the overall curves. This serves to further emphasize the inadequacy of the FWHM as a parameter to characterize collimators.

The line source response function due to scattered photons ($L_s(x)$) may be regarded as the inverse transform of the M.T.F. due to scattered photons ($\text{M.T.F.}_s(\nu)$). Due to the broad, flat nature of this latter line source response function the $\text{M.T.F.}_s(\nu)$ may be expected to approach zero very rapidly. Thus, above the value of the line frequency (ν) at which the $\text{M.T.F.}_s(\nu)$ is zero equation 4 reduces to:

$$MTF(\nu) = \frac{MTF_g(\nu)}{1+S} \qquad \ldots 5$$

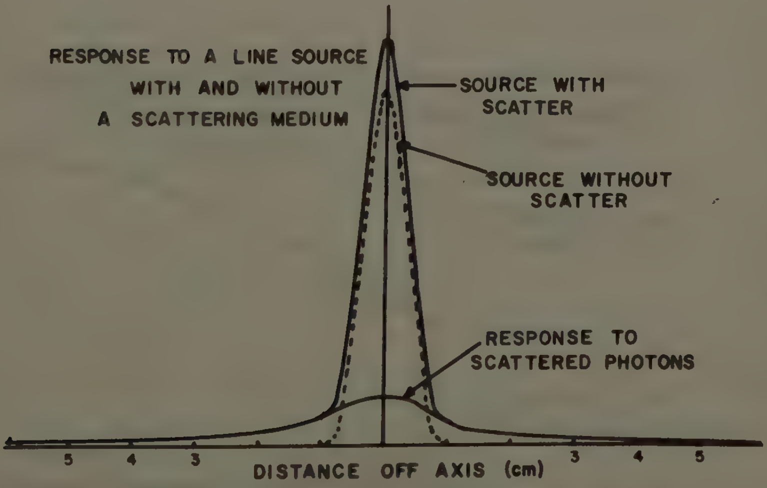


Fig. 8. A line source response function with and without scatter indicating the curve which may be assumed to be due to scattered photons alone.

---

* This figure reproduced in part from a paper given by Beck (37) at the III International Symposium on Medical Radioisotope Scintigraphy, I.A.E.A. 1968.

which can be re-written as:

$$S = \frac{MTF_g(\nu)}{MTF(\nu)} - 1 \qquad \ldots 6$$

where M.T.F.$_g(\nu)$ is the M.T.F. measured in air and M.T.F.(ν) is the M.T.F. measured in water above the limiting frequency (ν_{lim}) at which M.T.F.$_s(\nu)$ is zero.

Calculations such as those indicated in equation 6 have been performed on the M.T.F.'s obtained for a Nuclear Chicago collimator (#586) (31) at various depths in a scattering medium and the results are shown in Fig. 9. As is to be expected the scatter fraction increased from zero with no scatterer present to a plateau level of about 35% as the amount of scattering medium was increased.

Although this method of calculating the scatter fraction appears to be satisfactory in a number of cases, it does not work in all cases. The principal reason seems to be the uncertainty of when the M.T.F.$_s(\nu)$ falls to zero or if it does indeed fall to zero at all. Nevertheless, results so far tend to indicate that this method for calculation of scatter fraction does have some merit providing that the conditions which have to be met are fully recognized.

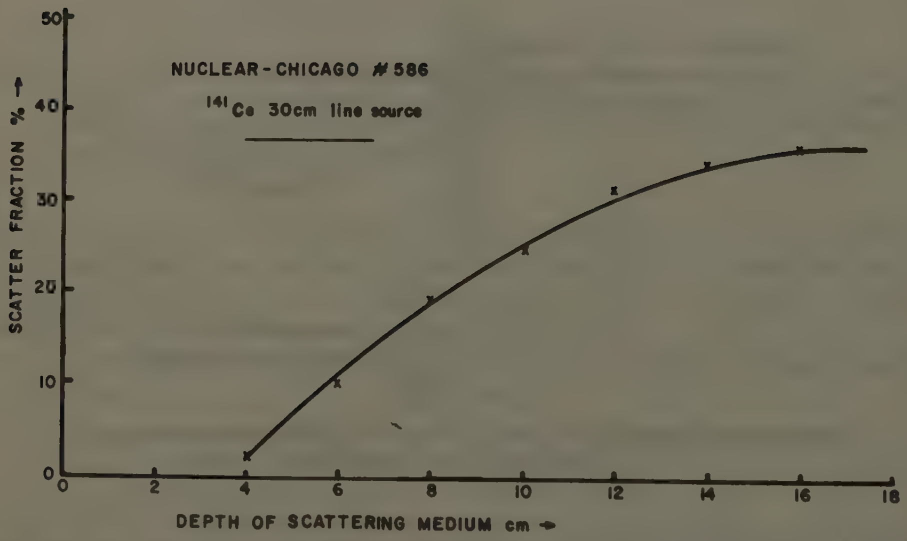


Fig. 9. A graph indicating the increasing scatter fraction with depth of scattering material for Nuclear-Chicago collimator #586. This curve was calculated from the collimator M.T.F.s' for each depth of scatter.

Penetration: If, instead of imposing a scattering medium between the source and collimator, we increase the energy of the source then the penetration factor can be measured using an analogous technique. Let us suppose that Fig. 10 represents the line source response functions first with a low energy source (negligible penetration) and then with a high energy source (penetration present), the difference between these two line source response functions may be regarded as the line source response function for penetrating photons ($L_p(x)$). As was the case for the response to scattered photons it is to be expected that this line source response function will also tend to be broad and flat and will give rise to an M.T.F.$_p(\nu)$ which has a low cut off or limiting frequency. Above this frequency and since scattering may be neglected we have:

$$MTF(\nu) = \frac{MTF_g(\nu)}{1+P} \qquad \ldots 7$$

where, in this case, M.T.F.(ν) is the M.T.F. with a high energy source and M.T.F.$_g(\nu)$ is the M.T.F. with a low energy source with a condition of no penetration being assumed.

Equation 7 may be re-written as:

$$P = \frac{MTF_g(\nu)}{MTF(\nu)} - 1 \qquad \ldots 8$$

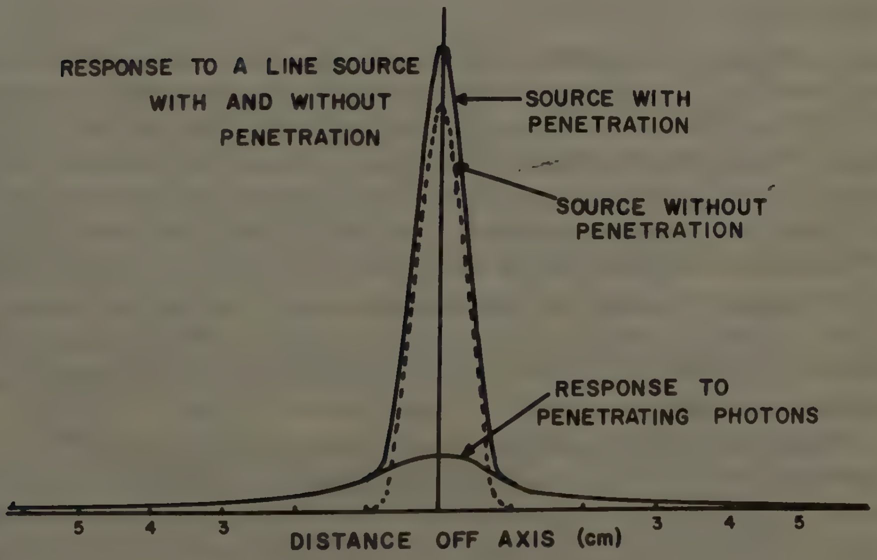


Fig. 10. A line source response function with and without penetration indicating the curve which may be assumed to be due to penetrating photons alone. No scatter is assumed to take place.

Calculations performed on information available on a Picker-Nuclear 2107A collimator (38) give values as high as 50% for the penetration fraction for I-131. This value agrees with that determined by Tsialis and Hine (39) using another method. The 2107A collimator is a 19 hole coarse focus collimator designed for I-131, but it has hexagonal holes whereas the original 2107 collimator had round holes and therefore thicker septa. It is interesting to note that this degree of septal penetration can be present yet in the case of no penetration the FWHM value is 1.6 cm and with penetration present this value is only increased to 1.9 cms. It is in the long tails of the line source response functions that the differences lie.

As was suggested above, when consideration was being given to the scatter fraction, this method of calculation is not without its drawbacks (Tsialis 40, Beck 41). Once again the inherent difficulty is to know when the $M.T.F._p(\nu)$ has fallen to zero so that equation 8 can be assumed to be valid. Further, although the line source response function for the scatter fraction may indeed be flat and long-tailed there is no guarantee that this will be the case for the penetration fraction. If a major portion of the penetration fraction is through the side shielding the line source response function may be long-tailed. If, however, as is more likely to be the case, the penetration is through the septa the $M.T.F._p(\nu)$ may not differ remarkably from the $M.T.F._g(\nu)$ and equation 8 cannot be used. Tsialis and Hine (39) have described a graphical method for the determination of the penetration fraction based on the line source response functions obtained with different energy sources.

Conclusion

An attempt has been made to present a review of the concept of M.T.F. and its applicability to radioisotope scanning systems. To date most of the emphasis has been upon performance analysis of the data acquisition components of the systems. Some work still remains to be done in this area for not all of the collimators presently available have been completely assessed and some data acquisition systems such as image intensifiers have yet to be analyzed. Some questions still remain to be answered on the subjects of penetration and scatter fractions and the methods of measurement of these parameters. This, however, will not preclude the application of M.T.F. concepts to other parts of the scanning system though, due to restrictions in linearity, referred to earlier, such application may be somewhat more difficult.

Further development may well see the innovation of the M.T.F. in two dimensions. This innovation has already been suggested by Harper et al. (42) in order to provide a theoretical prediction of the two dimensional image resulting from an object more closely resembling the type of lesion encountered in practice.

There can be little doubt that the concept of the M.T.F. has become well established as a useful tool in the assessment of scanner performance. Now that a standard form of measurement and presentation has been developed it will become much easier to intercompare different collimators and scanning systems.

REFERENCES

1. Cassen, B., Curtis, L., Reed, C., and Libby, R., Instrumentation for I-131 use in medical studies, *Nucleonics* **9,** (2), 1951 p. 46.
2. Mayneord, W. V., Turner, R. C., Newbery, S. P. and Hodt, H. J., A method of making visible the distribution of activity in a source of ionizing radiation, *Nature, Lond,* **168,** 1951, p. 76.
3. MacIntyre, W. J., Fedoruk, S. O., Harris, C. C., Kuhl, D. E., and Mallard, J. R., Sensitivity and resolution in radioisotope scanning. A report to the I.C.R.U., Symposium on Medical Radioisotope Scintigraphy, I.A.E.A., 1968.
4. Higgins, G. C. and Perrin, F. H., The evaluation of optical images, *Photo. Science and Eng.* **2,** 1958, p. 66.
5. Lamberts, R. H., Application of sine-wave techniques to image-forming systems, *J. Soc. Mot. Pic. and T.V. Eng.,* **71,** 1962, p. 635.
6. Altman, J. H. and Zweig, H. J., Effects of spread function on the storage of information on photographic emulsions, *Photo. Science and Eng.,* **7,** 1963, p. 173.
7. Higgins, G. C., Methods for engineering photographic systems, *Applied Optics,* **3,** 1964, p. 1.
8. Perrin, F. H., Methods of Appraising Photographic Systems. *J. Soc. Mot. Pic. and T.V. Eng.,* **69,** 1960, pp. 151 and 239.
9. Morgan, R. H., The frequency response function: a valuable means of expressing the informational recording capability of diagnostic x-ray systems, *Am. J. Roentgenol., Rad. Therapy & Nuclear Med.,* **88,** 1962, p. 175.
10. Morgan, R. H., Bates, L. M., Gopala Rao, V. V., Marinaro, A., The frequency response characteristics of X-ray films and screens, *Am. J. Roentgenol, Rad. Therapy, and Nuclear Med.,* **92,** 1964, p. 426.
11. Morgan, R. H., Visual perception in fluoroscopy and radiography, *Radiology,* **86,** 1966, p. 403.
12. Morgan, R. H., Threshold visual perception and relation to photon fluctuation and sine-wave response. Proc. 2nd Colloquium on Radiologic Instrumentation, University of Chicago, 1964, edited by R. D. Moseley. Thomas, Springfield, Ill., 1965.
13. Moseley, R. D. (Ed.): Proc. 2nd Colloquium on Radiologic Instrumentation, University of Chicago, 1964; Thomas, Springfield, Ill., 1965.
14. Rossmann, K., Measurement of the modulation transfer function of radiographic systems containing fluorescent screens. *Phys. Med. Biol.* **9,** 1964, p. 551.
15. Rossmann, K., The spatial frequency spectrum: a means for studying the quality of radiographic imaging systems, *Radiology,* **90,** 1968, p. 1.
16. Doi, K., Optical transfer functions of the focal spot of x-ray tubes, *Am. J. Roentgenol, Rad. Therapy, and Nuclear Med.,* **94,** 1965, p. 712.
17. Doi, K. and Sakamoto, H., Optical transfer treatment of image transfer in radiography, Progress in Radiology, Vol. II, p. 1682, Excerpta Medica Foundation, 1967.
18. Beck, R. N., A theory of radioisotope scanning systems, Medical Radioisotope Scanning, Vol. 1, I.A.E.A., Vienna, 1964.
19. Gottschalk, A., and Beck, R. N. (Eds.), Fundamental Problems in Scanning, Thomas, Springfield, Ill., 1968.
20. Mozley, J., Quantitative comparison of radioisotope scanners and cameras in brain tumor imaging., *J. Nuc. Med.,* **10,** 1969, p. 359.
21. Cradduck, T. D., Assessing the performance of radioisotope scanners: data acquisition, *J. Nuc. Med.,* **9,** 1968, p. 210.
22. Gregg, E. C., Pedagogical note on M.T.F.'s, *Investigative Radiology,* **1,** 1966, p. 418.
23. Hine, G. J., Evaluation of focussed collimator performance II: Digital recording of line-source response, *Int. J. Appl. Rad. Isotopes* **18,** 1967, p. 815.
24. Frey, H. S., Evaluation of photoscanners, *Investigative Radiology,* **1,** 1966, p. 314.
25. Gopala Rao, U. V., and Wagner, H. N., Effect of an analogue ratemeter on the modulation transfer function in radioisotope scanning, *Radiology,* **88,** 1967, p. 504.
26. Cradduck, T. D., Fedoruk, S. O., and Reid, W. B., A new method of assessing the performance of scintillation cameras and scanners, *Phys. Med. Biol.,* **11,** 1966, p. 423.
27. Cradduck, T. D., A simple method of measuring the line source response function and modulation transfer function of a scintillation camera, *J. Nuc. Med.,* **10,** 1969, p. 397.

28. Keller, E. L., and Coltman, J. W., Modulation transfer and scintillation limitations in gamma-ray imaging, Westinghouse Research Report #67-1C2-CAMRA-P2, 1967.
29. Hine, G. J. and Tsialas, S. P., Evaluation of focused collimator performance: III Three- and five-inch diameter collimators for I-131 and Tc-99m. Symposium on Medical Radioisotope Scintigraphy, I.A.E.A. 1968.
30. Ross, R. D. and Cradduck, T. D., Line-source response function measurements and modulation transfer function calculations performed on some Nuclear-Chicago and Ohio-Nuclear collimators. Symposium on Medical Radiosotope Scintigraphy, I.A.E.A. 1968.
31. Ross, R. D., An experimental evaluation of some commercial collimators used in rectilinear scanners. Masters Thesis, University of Manitoba, 1968.
32. Christie, J. H. and MacIntyre, W. J., A comparision of line spread functions of five- and eight-inch collimators with three-inch and tapered bore collimators. Symposium on Medical Radioisotope Scintigraphy, I.A.E.A., 1968.
33. Iio, M., Evaluation of focused collimator performance: Collimator made by three Japanese scanner manufacturers. Symposium on Medical Radioisotope Scintigraphy, I.A.E.A., 1968.
34. Tsialas, S. P. and Lewis, B., Division of Life Sciences and Computer Services Section, I.A.E.A., Private communication, 1969.
35. Mallard, J. R., A proposal to standardize the evaluation of all counters, The Design of Counting Systems for Dynamic Studies and Uptake Measurements, British Institute of Radiology, 1968.
36. Sanderson, G. K., Erroneous perturbations of the modulation transfer function derived from the line spread function, *Phys. Med. Biol.*, **13**, 1968, p. 661.
37. Beck, R. N., Schuh, M. W., Cohen, T. D., and Lembanes, N., Effects of scattered radiation on scintillation detector response. Symposium on Medical Radioisotope Scintigraphy, I.A.E.A., 1968.
38. Hine, G. J., Private communication, 1967.
39. Tsialas, S. P. and Hine, G., Collimator characteristics for radioisotope scanning, Submitted to J. Nuc. Med., February, 1969.
40. Tsialas, S., Private communication, 1969.
41. Beck, R., Private communication, 1969.
42. Harper, P. V., Beck, R. N., and Reddy, M., Two dimensional image analysis in radioisotope scanning systems, *J. Nuc. Med.*, **10**, 1969, p. 340.

AN IMAGE INTENSIFIER ISOTOPE CAMERA CAPABLE OF PULSE HEIGHT DISCRIMINATION

M. M. Ter-Pogossian

A number of investigators have explored the possibility of incorporating an image intensifier tube into the design of an isotope camera. Mortimer *et al.* (1) and Kellershohn and Pellerin (2) attempted to visualize radioactive sources by means of X-ray image intensifier tubes. These early designs, however, did not lead to the development of practical clinical isotope cameras principally because the unfavorable signal-to-noise ratio of the intensifiers used at that time allowed only the visualization of radiation sources too intense for diagnostic clinical applications.

The development of high signal-to-noise ratio image intensifier tubes specifically designed for this purpose permitted the design of an isotope camera capable of recording single photon interactions (Ter-Pogossian *et al.* (3)). This design, after numerous improvements, evolved into the construction by the "Picker" Company of a clinically useful instrument, the "Magnacamera". The detector of this instrument contains a nuclear image intensifier tube specifically designed for the purpose with an input screen composed of activated crystalline cesium iodide approximately 1.5 millimeters thick and about 8 inches in diameter. The output screen of the image intensifier tube is optically coupled to an intensifier orthicon television camera tube. The video signal generated by the television camera tube is electronically processed. This instrument has been described before (Ter-Pogossian (4), and the present text will include only a discussion of the advantages and drawbacks of this instrument, in addition to alterations of this design.

The image intensifier tube acts as a light pre-amplifier to raise the luminance of the light image formed by the X- or gamma-ray photons interacting in the cesium iodide input scren. The luminance of the image formed on the output screen of the tube is further raised by the intensifier orthicon television tube and converted into a video signal. The image intensifier tube thus functions as a low-noise pre-amplifier, and the intensifier orthicon functions to further intensify the luminance of the image and to convert it into a video signal. The conversion of the optical image into a video signal is particularly useful because it allows its recording on video tape, and the electronic manipulation of the video pulse image for quantization and pulse height analysis. A facet of this design, the importance of which deserves explanation is the fact that the crystal which converts into X- or gamma-ray image into a light image is placed within the vacuum envelope of the image intensifier tube in intimate contact with its photocathode. This design exhibits a number of drawbacks: 1. the introduction of the crystal into the envelope of the tube complicates the construction of the image intensified; 2. the size of the input screen is limited to the diameter of the intensifier tube, and large diameter intensifier tube are difficult to build; and 3. it is impossible to change the input screen should it appear desirable to trade resolution for efficiency. These drawbacks suggest the placement of the input screen (crystal) outside the vacuum envelope of the image intensifier with an optical coupling between the latter and the photocathode of the tube. Although this approach has indeed been suggested and tested by a number of investigators, it has not yet led to the development of a practical isotope camera. The difficulty in this approach is in achieving a sufficiently efficient optical coupling between the input screen and the photocathode of the image intensifier to prevent the quantum sink of the system from dropping below a value where noise interferes with the visualization of single X- or gamma-ray photons and with meaningful pulse height analysis. Various optical couplings can be used, and some have been tested for this purpose. They fall into one of three categories: 1. lens transmission or reflection optics, 2. fiber optics coupling between the input screen and the glass surface of the image intensifier tube, and 3. utilization of an image intensifier tube with a fiber optics input, allowing direct fiber optics coupling between the input screen and the photocathode of the tube. The first approach results, in general, in a prohibitively high loss of photons unless large, bulky, and expensive optics are used. The second approach is accompanied by a loss of resolution due to the diffusion of light in the glass envelope of the image intensifier tube. In spite of this weakness, however, this second approach appears to offer promise. The third approach is by far the most attractive; but unfortunately, at this time large image intensifier tubes with fiber optics input are not readily available. Under the circumstances, a design such as shown

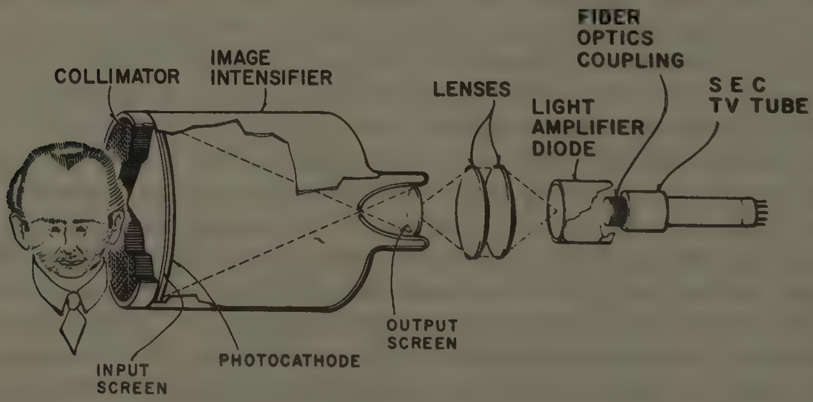


Fig. 1. Diagram of the components of the detector system of the SEC-image intensifier isotope camera (Magnacamera).

in Figure 1, which provides an excellent optical coupling between an input screen and the photocathode is a practical compromise until the availability of fiber optics input image intensifiers would allow a more flexible arrangement.

The described camera exhibits a number of desirable features but also suffers from some serious disadvantages. First, the inherent resolution of the Magnacamera is, for all practical purposes, limited only by the thickness of the crystal used. With a 1.5 millimeter thick crystal the system is capable of resolving approximately 4 lines per centimeter. With such a crystal the efficiency of the camera for the detection of X- and gamma-ray photons is shown in Table I. It is apparent that this system is suitable only for the detection of low and medium energy radiation. Above approximately 300 kiloelectron volts the efficiency of the system drops below 10 percent. This efficiency can be raised at the cost of resolution by using a thicker input screen crystal.

TABLE I

Photon Energy (keV)	Approximate Detection Efficiency %
40	45%
80	80% (−)
150	40% (+)
300	10%

The radioactivity, per unit area, of the organ visualized by the camera can be "quantified" by integrating the television signal within a preset area. Thus, quantitative dynamic studies can be carried out.

One of the great advantages of the design of this camera is the fact that, for all practical purposes, any flux of X- or gamma-rays can be processed. Thus, very fast dynamic studies are possible when large amounts of a radioactive tracer can be safely administered to the patient.

Perhaps the most serious disadvantage of this design is the fact that this camera is incapable of pulse height analysis because the orthicon tube supplies broad, unsharp pulses which do not lend themselves to electronic pulse height discrimination. This drawback was corrected to a great degree by replacing the orthicon television tube by an SEC-vidicon tube (Figures 1 & 2). The possible usefulness of the incorporation of an SEC-vidicon tube into the design of an isotope camera was suggested by Sternglass and Keller (5). This tube possesses a number of characteristics which are particularly suited for the described purpose. In particular, it converts flashes of light nearly linearly into sharp pulses which can be height analysed. Furthermore, the SEC-vidicon possesses a broad dynamic range greater than 100 to 1 (5) in comparison to a range of approximately 10 to 1 for the orthicon tubes previously used. Unfortunately, the light sensitivity of the SEC-vidicon is considerable lower than of an image intensifier orthicon tube, and the incorporation of the former in the described isotope camera requires additional

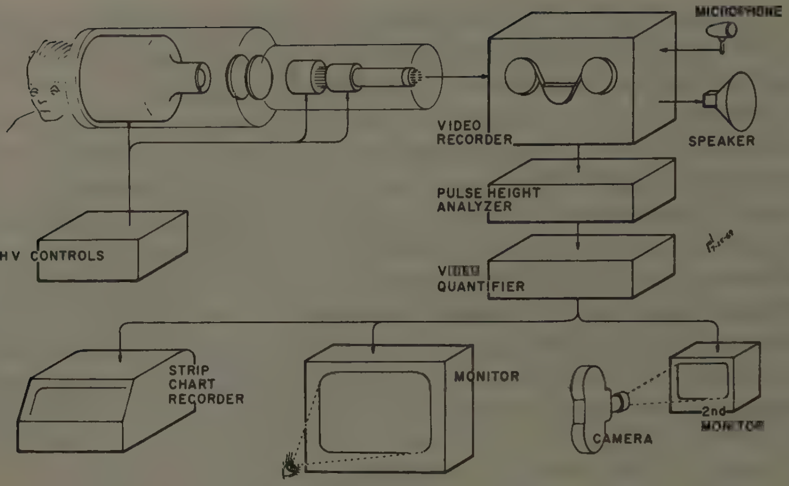

MICROPHONE
VIDEO RECORDER
SPEAKER
PULSE HEIGHT ANALYZER
HV CONTROLS
VIDEO QUANTIFIER
STRIP CHART RECORDER
MONITOR
CAMERA
2nd MONITOR

Fig. 2. Block diagram of the components of the Magnacamera.

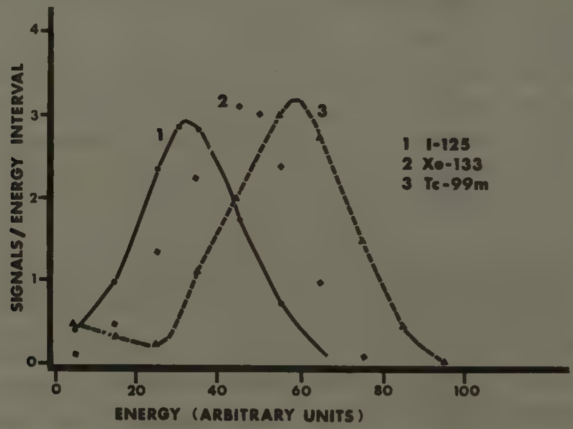


Fig. 3. Family of curves showing pulse height resolution versus gamma-ray photon energy impinging upon the input screen.

light amplification between the image intensifier tube and the television pick-up tube. This has been achieved by the insertion of a light diode as shown in Figure 1.

The replacement of the orthicon television tube in the described design by the combination light amplifier dynode-SEC-vidicon tube corrected one of the most important disadvantages of this design by allowing pulse height analysis. This pulse height analysis is not as precise as achieved with scintillation counters as shown in Figure 3 because of the degradation of the linear relationship between the energy of the absorbed X- or gamma-ray photon in the input screen and the video signal supplied by the SEC-tube. Nevertheless, this pulse height analysis of the video signal allows the rejection of the major portion of the scattered radiation in most clinical studies and improves considerably the quality of the image obtained (Figure 4).

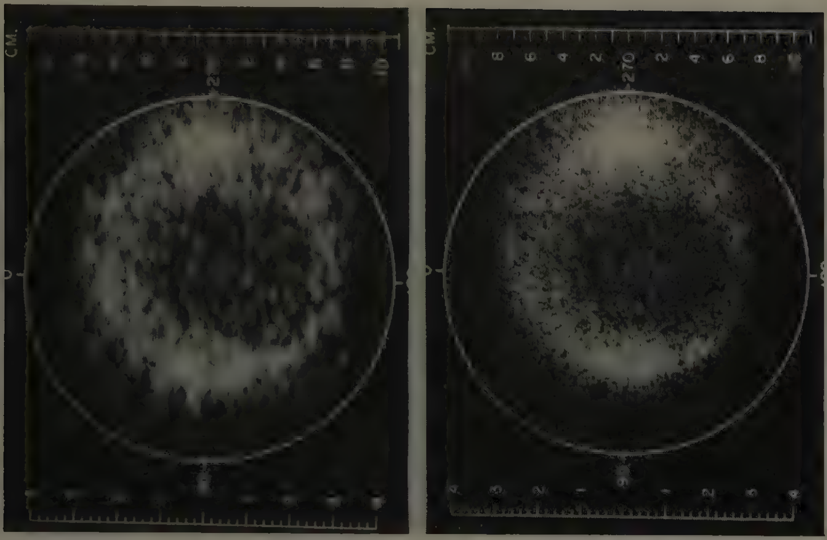


Fig. 4. Demonstration of the improvement achieved through the use of pulse height analysis. AP view of the skull of a patient obtained with the camera subsequent to the intravenous injection of 10 millicuries of technetium 99-m pertechnetate. The picture was obtained approximately ½ hour after the injection of the activity and the integration time was 4 minutes. The (right) picture obtained from the same video tape as (left) for the same period of time but with pulse height analysis.

The above described camera exhibits the following distinct advantages:

Its inherent resolution is so high that for all practical purposes the quality of the image obtained is limited only by the collimator used and by the statistical fluctuations of the photons. In many instances when the recorded image is formed by a large photon flux, this feature yields in a definite improvement of the quality of the image.

This camera is capable of processing very high photon fluxes and thus, for all practical purposes, its use is not limited by the dead time of electronic circuitry, unless pulse height analysis is used.

The camera is endowed with a particularly broad recording dynamic range. It is capable of recording in the analog mode extremely high fluxes of photons even though pulse pile-up does occur. For lower fluxes the information is processed as separate analyzed pulses. The latter feature allows the imaging of the passage of a concentrated bolus of activity through an organ (Figure 5) followed by repeated imaging of the same organ after considerable dilution of the activity.

In this camera pulses are first recorded on tape, and pulse height discrimination is carried out from the recorded tape. This allows the most suitable level of discrimination to be established after recording (sometimes by trial and error) without the disadvantage of the possible loss of useful information as when pulse height analysis is preset for the examination.

This camera is presently undergoing clinical evaluation.

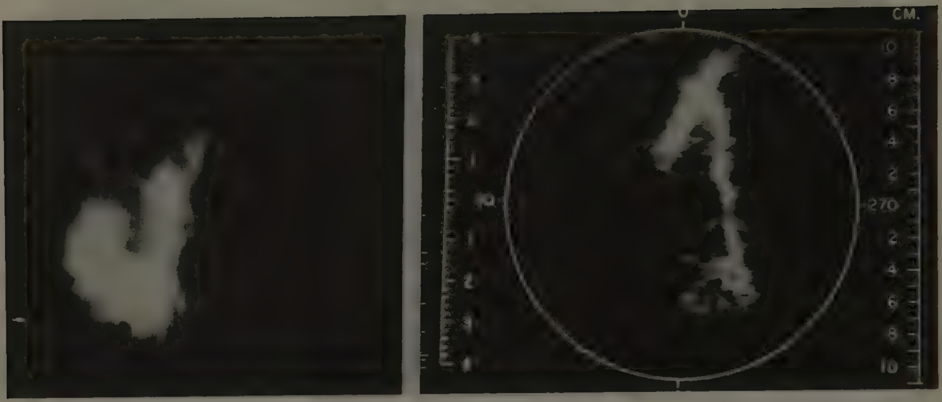


Fig. 5. (A & B). Two pictures of the heart obtained subsequent to the intravenous injection of 10 millicuries of technetium-99m pertechnetate. Integration time 0.25 seconds. Figure (5 A) normal heart. Figure (5 B) obstruction of the vena cava and collateral circulation leading to the heart.

REFERENCES

1. R. K. Mortimer, H. O. Anger and C. A. Tobias, The Gamma-Ray Pinhole Camera with Image Amplifier. University of California Radiation Laboratory Report U. C. R. L. -2524 (1954), also in "Convention Record of the Institute of Radio Engineers" (1954) - National Convention, The Institute of Radio Engineers, New York, Part 9, pp. 2-5 (1954).
2. C. Kellershon and P. Pellerin, Sur la Possibilite d'Utiliser un Amplification d'Image Pour Mettre en Evidence la Localization et la Distribution d'un Corps Radioactif. Comptes Rendus des Seances de la Societe de Biologie CXLIX, 533 (1955)
3. M. Ter-Pogossian, J. Kastner and T. B. Vest, Auto fluorography of the Thyroid Gland by Means of Image Amplification. Presented at the Meeting of the RSNA, 1962, in Radiology 81, #6 984 (1963).
4. M. Ter-Pogossian, W. F. Niklas, J. Ball and J. O. Eichling, An Image Tube Scintillation Camera for Use with Radioactive Isotopes Emitting Low Energy Photons. *Radiology* **86,** #3, 463 (1966).
5. E. J. Sternglass and E. L. Keller, High-Resolution Electronic Gamma-Ray Imaging, in Advances in Dynamic Scanning. Edited by Y. Wang, Charles C. Thomas, Springfield, Ill., (1968).

ULTRASONIC IMAGING WITH IMAGE CAMERAS, HOLOGRAPHY AND COLOUR-PHASE REPRESENTATION

Ward J. Collis
C. N. Smyth

INTRODUCTION

To construct a complete image the angular distributtion of intensity, frequency and phase of the scattered radiation must be known. Intensity and frequency, or amplitude and frequency, are comparatively easy to determine but relative phases are not (1). Construction of images by lenses avoids the problem through directly using the waves scattered from the object and preserving the correct phase relationships. Lipson has so aptly said also that the whole basis of the design of optical instruments lies in the preservation of the relative phases of the scattered radiation.

So many objects that are opaque to visible light are translucent to sound, or trans-sonic, that there are evident advantages in attempting to use such radiation to visualize the internal structure of the body, or the internal imperfections of engineering components or materials.

Ultrasound in the 'Mega-Hertz' region in liquid media provides waves of 1mm wavelength or less and these are admirably suited to the delineation of structures large enough to be seen by the naked eye. The resolution is determined by the aperture of the system and the wavelength used and is $\lambda/\sin a$, where a is the angular aperture and λ is wavelength, irrespective of the method. For an aperture of 22° and a frequency of 4.4 MHz in water the wavelength is 0.34mm and the resolution 0.9mm, figures which there will be occasion to refer to again later. Most of the experimental work referred to in this paper was completed at this frequency which is the sub-carrier frequency for UK colour television.

The problem of ultrasound imaging is, however, beset with difficulty, and the more consequential hazards are listed here lest the reader should come to too hasty a conclusion regarding the usefulness of the method or plan new research of an unrewarding character.

---

* Department of Electrical Engineering, Northwestern University, Evanston, Illinois.

Many techniques are available by which ultrasonic waves can be employed to produce a visual representation of the internal and external boundaries of opaque objects, yet only one of these methods, the pulse-echo Sonar direction ranging, has found a wide-spread usefulness in medicine and in engineering (2). The reasons for this success over other methods lies largely in the relative unimportance of phase information which contributes almost unobtrusively to the directional properties of probes and echoes. The receiving equipment and viewing devices are only concerned with amplitude. Furthermore, the pulse-echo system has an inbuilt aptitude to distinguish interfaces at depth unconfused by surface glare or reflections.

APPLICATIONS OF ULTRASONIC IMAGING

Changes in elastic continuity in supposedly homogeneous materials although not evident on X-ray inspection are readily perceived by ultrasonic techniques. Defects in castings, forgings, failures of bonding, welding, electroplating gave ultrasound an undisputed area of usefulness in nondestructive testing. The detection of air-bubbles in rocket propellants is another example of established usefulness. For these nondestructive techniques, variations in the intensity or transmission time-delay of Sonar reflections provide the information required. Inanimate engineering components can be patiently scanned by automatic devices using focused or collimated beams because time is not important. In mass production units, speed of scanning may be consequential and here there is room for development of multiple scan and differential detecting methods.

For complicated objects or those with multiple internal reflections the pulse-echo method can be confusing and it would be advantageous if a more nearly optical parallel could be achieved and the specimen rendered visibly translucent, as though it were made of lucite or perspex. Such visualization is not impossible but internal interfaces might overlie one another and curved surfaces act as lenses distorting those parts it was most desired to see.

In the medical field Sonar sections of the body have become indispensible for accurate diagnosis of gynecological tumours, facets of pregnancy, lumps in the liver and breast, central neurological disorders, tumours of the eye and orbit and other applications.

It is therefore very desirable before developing new and perhaps very expensive methodology to define the engineering and medical problems which at present escape our grasp. In the medical area the development of the Doppler principle movement detector and its application to heart action and blood flow measurement has in a simple way removed much of the need for visualization. It would be interesting to observe heat valves in action but the surgical judgment would still probably rest with measurements of their dynamic action and flow measurements, neither of which calls for visual observation.

Temporizing then, fixed or static shapes may (if time permits) be explored by scanning, moving parts may be explored by rapid scanning or direct imaging and on occasion the devices could profitably be arranged to visualize only moving edges and to disregard stationary interfaces (3). The flow of oil in machinery, the welding process, and chemical treatments in the inanimate world might become useful applications. In biology or medicine, blood flow, heart action, the closing of arteries on injury or after parturition, the separation of the placenta, movements of the bowel, the intactness of internal sutures, the position of catheters, the alignment of bones, the shape, movements and habits of the fetus—sex, thumb sucking, cord biting, swallowing, micturition, eye movements—if made visible might add immeasurably to our store of knowledge and on occasion assist medical practice directly. Ultrasonics has little place for visualizing the external shape of objects, thick or thin, in light or dark; this function is the prerogative of the eye and hand, of the senses of sight and feeling. Ultrasonics might however serve to transduce similar information to the input circuits of a computer, or to delineate the contours of solids immersed in murky water or foam.

When solids are examined by ultrasound the surface, even when the object is immersed in a suitable medium, provides the strongest interface reflection, or refraction if the contour is curved. For transmitted sound the surfaces may not obscure the outline of internal interfaces, but, when the sound or ultrasound is reflected from the object to the viewing device surface glare may saturate the dynamic range of the transducer and prevent visualization of internal parts or the far surface. This disadvantage is overcome by the use of pulses of ultrasound from the transmitter and gating the receiver to function only with echoes from a chosen depth. Tomograms are then produced rather than complete three dimensional photographs.

METHODS AND COMPLICATIONS

Pulse-echo Sonar has been mentioned and is well documented (2). Its capabilities are immense for delineating shapes in two dimensions, demonstrating the extent of movements and making dimension comparisons between two sides of the body, or between good and defective components. Its disadvantage is that resolution is limited by beam size, and the scanning process is slow, pictures in the human subject being affected even by respiration. On the other hand, it is very sensitive and echoes can be obtained throughout the depth of the body and from 100 meters in metal even at 1 MHz. At lower frequencies the method works for vast distances in sea water. At 10 MHz the penetration into the body may be limited to a few cms depending on the nature of the tissue. Bone is extremely dispersing and lung tissue, unless collapsed, may appear opaque above 1 MHz.

Visualizing techniques may be divided into the categories of scanning and image converting methods on one hand, and diffraction pattern and hologram reproduction procedures on the other. The diffraction pattern group includes picture formation by shadow or acoustic lens methods. In the hologram group the amplitude and phase of the ultrasound field in a convenient plane is recorded for visualization by optical methods (often with laser illumination (4) and optical magnification to restore the image size) after the acoustic hologram has been dimensionally reduced to be compatible with optical wavelengths (vide infra.). All of these four methods can be operated with pulsed ultrasound for tomogram production.

In the choice of a method, the decision between direct imaging or scanning, or a combination of both (for the image may be stored before being read out by scanning) is usually a matter of deciding how long the picture build-up may take. If the object is insonated continuously then the scanning process will take from 1/50th to 1/1000th second. With pulses however the scanning spot must wait at each picture element for the arrival of the wave packet. These cannot usually be more frequent than 1000 Hz if internal reflections and camera—object reflections are to decay before the pulse is repeated. This method therefore increases the scanning time to the order of 1000 seconds. The time can be reduced by storing the whole image from a single pulse and reading the image by scanning subsequently in a scanning frame time. A tube to do this is described in the Section on 'Experimental Procedures'.

Holography (4), which may be accomplished by scanning or imaging direct, is theoretically able to offer a number of advantages: the supposedly difficult ultrasound lens may be replaced by optical lenses; the resulting picture can be stereoscopic or '3D' (although this has not yet been achieved with sound because of the wavelength disparity); one hologram can represent a number of co-planar tomograms which can be focused on separately.

It is not practical to make a 300:1 or more reduction without affecting image quality. Dimensional reduction is either modest or, in the case of fast systems, not performed at all.

Holograms will show depth of focus but not 3D effect owing to limited aperture size, in terms of wavelength. On the other hand, the holograph method is not 'on line' (or in real time). The photographic processing delay is currently of the order of weeks except in long wave examples (with under water experiments or airborne sound or radio waves) where computer processing of the data is nearly instantaneous. On line holography from a liquid gas interface has been demonstrated experimentally.

Acoustic shadow or reflection methods where the camera-object distance is small, produce pictures which are recognizable and often useful. Most engineering and biological surfaces are specularly reflecting at millimeter wavelengths and this may assist shadow-reflection techniques once the character of the image is appreciated by the observer. A flat speculum produces an image of the source transmitter, curved surfaces produce distorted real or virtual images of the source. For this reason it is sometimes useful to use a source of recognizable form and rotation e.g. a triangular or arrowhead shape.

The implication of much recent work with ultrasonic holography is that it circumvents the need for an unobtainable acoustic lens. It is not difficult to design and make acoustic lenses free from axial distortion but they are subject to errors off-axis arising from acoustic absorption in the thicker parts, from internal reflections, and the difficulty of blooming to reduce surface reflection, especially with the large curvatures usually involved. In our experimental work to date, the results with lenses have been superior to those with holography.

Finally, in this section, must be mentioned the additional requirement of holography that the object and intervening medium must be constant or stationary during the making of the hologram so that the reproduction of the image can reproduce the correct phase relationships over the whole image. A wavelength of sound at 4.4 MHz is about 0.3mm. The object and intervening fluid should remain stationary and have uniform temperature so that coherence is not lost in the coupling medium. Assuming that some parts of the body move at 100 cms/second, it follows that the hologram needs to be made in less than 100 microseconds. If the medium is turbulent it may not be possible to reproduce the hologram unless a separate hologram representing the turbulence alone can be produced simultaneously. The situation is analogous to the use of optical holography for seeing through frosted glass which has so intrigued the experts in this subject. It was solved by collecting the glass for subsequent rectification of the hologram!

When all these factors are considered the use of ultrasonic lenses may have much to recommend it.

Experimental Procedures

1. The Ultrasound Image Camera (3)
2. The Ultrasound Image Storage Camera (3)
3. Phase Sensitive Colour Pictures (9)
4. Holography With Electronic Scanning (10)
5. Holography by Direct Image Methods
6. Holography With Mechanical Scanning (11).

The authors have had and are gaining experience with the first five of these methods in a programme which has been running since 1954. The last method, the production of holograms by mechanical scanning with a search probe and electronic processing of the signal is currently under feasibility study at E.M.I. (Wells) Ltd. and at the Electronics Section of the Atomic Energy Research Establishment, Harwell, England.

Descriptions of these methods have been published by the authors and others as the references denote, and the present brief presentation is planned to compare the usefulness and feasibility of the various techniques. Since the recent stage of the work commenced, and indeed since this publication was first prepared, there have been various national and international conferences on holography (12).

The Ultrasound Image Camera

The ultrasound image camera or ultrasonic microscope as it was first titled by the inventor Sokolov in 1936 (5) (Figure 1) is a modification of the Zworykin television camera in which the photo-electric mosaic is replaced by a bare piezoelectric quartz plate resonant in the half wave mode to the frequency of the investigation; in Sokolov's example 3000 MHz, in our work 4.4 MHz. The camera was

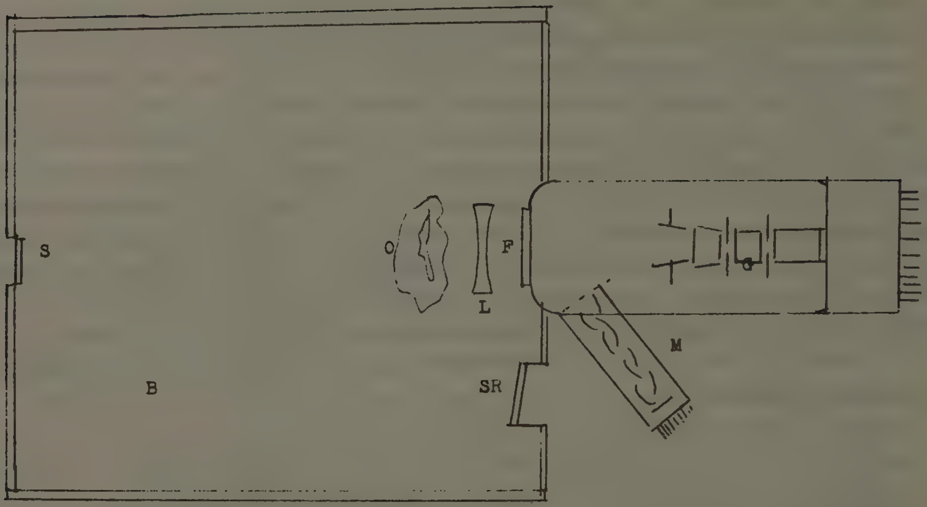


Fig. 1. Sokolov Ultrasonic camera as improved by Jacobs.

S	ultrasonic source for transmission through object O
SR	position for reflecting source to insonate object
L	acoustic lens
F	camera resonant quartz face plate, uncoated
G	electron gun with electrostatic scanning
M	electron multiplier and position for introduction of reference signal during holographic operation.

built by the authors as a sealed-off tube in 1955 and made to work with low velocity scanning so that it would differentiate phase (13) and also moving from steady diffraction patterns (3). Subsequently, a modification of the Sokolov tube was made on a continuously pumped system by Fritag and Martin (6), and later Jacobs much improved the design by the addition of a multi-stage electron multiplier and the successful sealing of large quartz plates to glass at high temperature (7). These tubes had a large dynamic range and the performance of the systems was limited by the dynamic range of the television viewing arrangements. Saturation of amplifiers often increased the apparent resolution.

Published results at television standards of scanning gave sensitivities in excess of 0.1 μ watts/cm^2 and resolution about 0.5mm. In an image, for example, the gross structure of the finger and parts of the hand could be seen. Jacobs showed movement of tissues within the heart of a cat. Other workers were not so successful with the intact animal. In the rat we had difficulty in distinguishing heart and diaphragmatic movements. Surface reflections obscured the image and reflected gated pulse tomograms made the scanning too slow to show moving tissues in the bodies of man or animals.

The Storage Camera

To overcome the difficulties found with the Sokolov (5) and low voltage scanning cameras (3), a camera was designed, constructed, and is currently under test in which the sound field pattern on the transducing face can be fixed instantaneously and subsequently read off by scanning. The plan of this tube is shown in Figure 2. It has been discussed previously (8), but only recently satisfactorily engineered by one of our colleagues*. The mode of action is described in the caption of the picture. No pictures representative of its performance are yet available. It has a 1 inch diameter face-plate. Additional grids are necessary if it is to be used for holography with an artificial oblique reference signal.

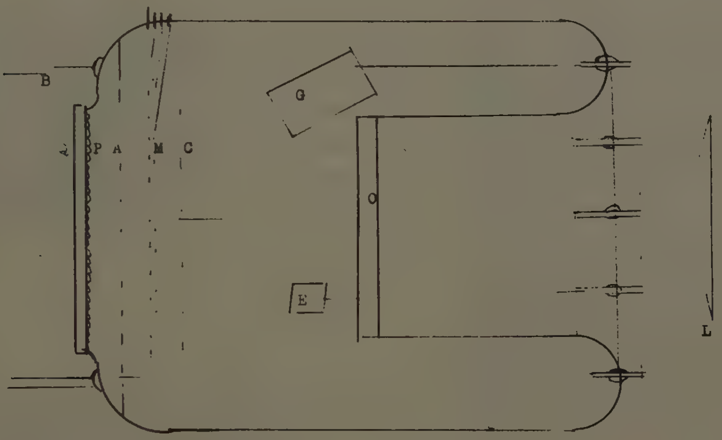


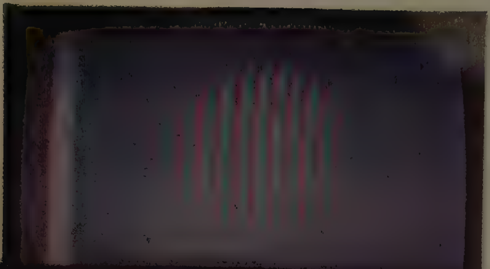
Fig. 2. Light beam scanned ultrasonic camera.

The image can be fixed on the face plate by an impulse from the flood gun, and the charge pattern which mimics the ultrasound pattern at the time of the impulse, can be read out by light beam scanning. For holography the sound pattern must be read out by scanning while the ultrasound is on, and the flood gun used for recharging only. For holography additional multiplier grids are added to simulate an oblique reference beam

Symbols: B waterbath
Q quartz face plate transducer
p photoelectric mosaic
A accelerator grid
M multiplier grid matrix
C collector
G flood gun
E evaporator for processing
O optical window
L lens
S cathode ray tube scanned light source.

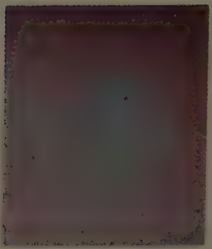
---

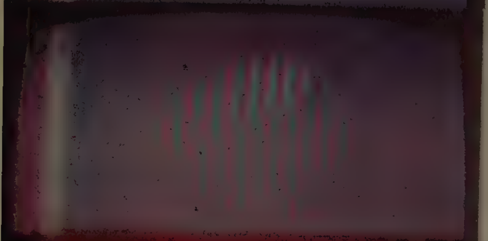
* J. E. Ewins



3a

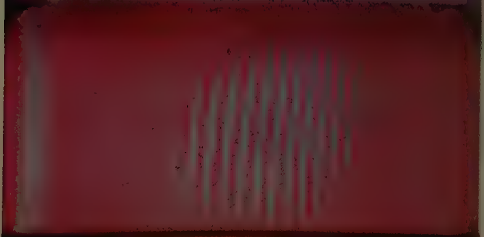
9d



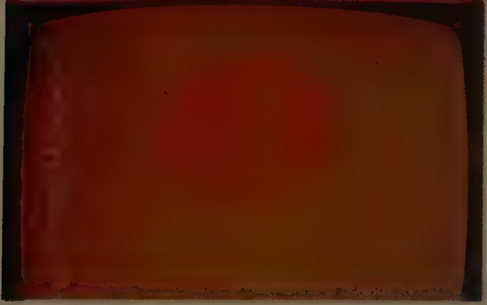


3b

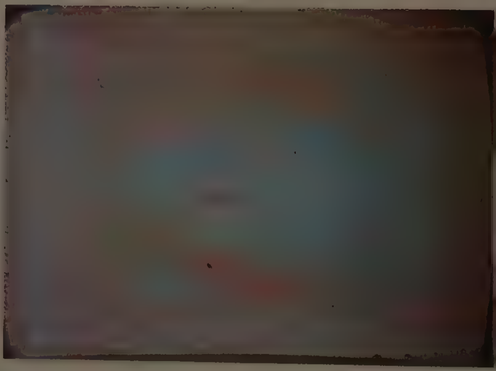
10c



3c

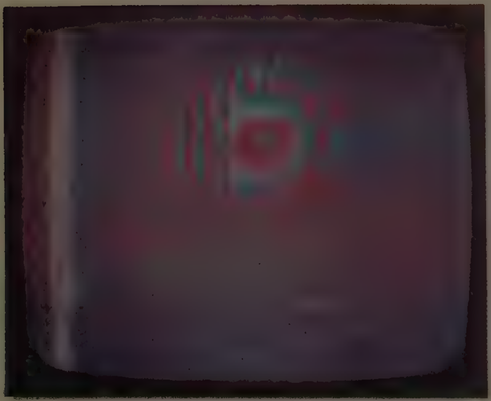


3d

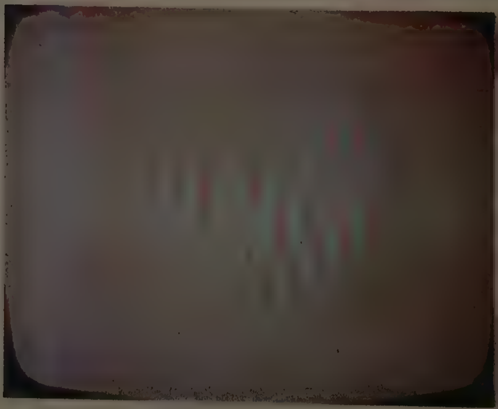


11a

4



11b



Representation of Phase by Colour

Jacobs (9) first employed conventional colour television receivers scanned at conventional television standards to denote the phase and intensity of the ultrasound diffraction patterns on the camera face. This is technically straight forward when the ultrasound source is synchronous with the colour sub-carrier frequency in the television receiver. In the English 'compatible' system the frequency is 4.4 MHz. Also, for European television receivers which have colour reversal in alternate frame scans, minor circuit modifications are necessary. The resulting phase contrast images shown in Figures 3 and 4 (see color plate p. 192) are colour "holograms" since at every point the amplitude and phase of the signal are separately defined. Whether such holograms contain more information than the conventional monochrome ones depends on the laser light combinations available for simultaneous viewing.

The phase colour method does however often make images more recognizable as the pictures of Figures 10 & 11 show. (see color plate p. 192.)

Colour phase patterns of images formed by lenses assist the experimental adjustment of lenses and may make lens pictures comparable with holographic possibilities while retaining the 'on line' facility. Figure 4.

Fig. 3. a) Source Hologram; b) Edge of aluminum sheet hologram, $\lambda/3$ phase delay; c) fine wire 1/2 mm. hologram; d) ultrasonic shadow of wire.

Fig. 4. Focusing action of perspex (lucite) lens on oblique ultrasonic beam—an approach to lens design.

Fig. 9d. Shadow of cross-pattern, phase color picture

Fig. 10 c. Color phase picture of finger shown in holograms 10a and 10b.

Fig. 11. Perspex (lucite) block with center hole. a. Phase color picture; b. Phase color hologram, 2nd order hologram.

Holography With Electronic Scanning

Figure 5 shows the basic holographic amplitude-phase recording method using an off axis reference signal. The amplitude is conveyed by the peak voltage or brightness of the visual interference fringes and the phase by their position in the holograph plane.

Employing quartz plates as the hologram transducer, this double acoustic source method is not satisfactory because the piezo-electric effect is an adding and not a multiplying effect. Additionally, zero order holograms lead to confusion since the various waves from the object add together, and cannot be distinguished from the sums of the object and reference waves. The Fourier transform of the addition products leads to an image with little information and confused by glare from the source (Figure 7). On the other hand, the product of the object and reference signals, if these can be mixed in a pure multiplier with an oblique reference beam, contains all the information and there are no higher terms to be included since the

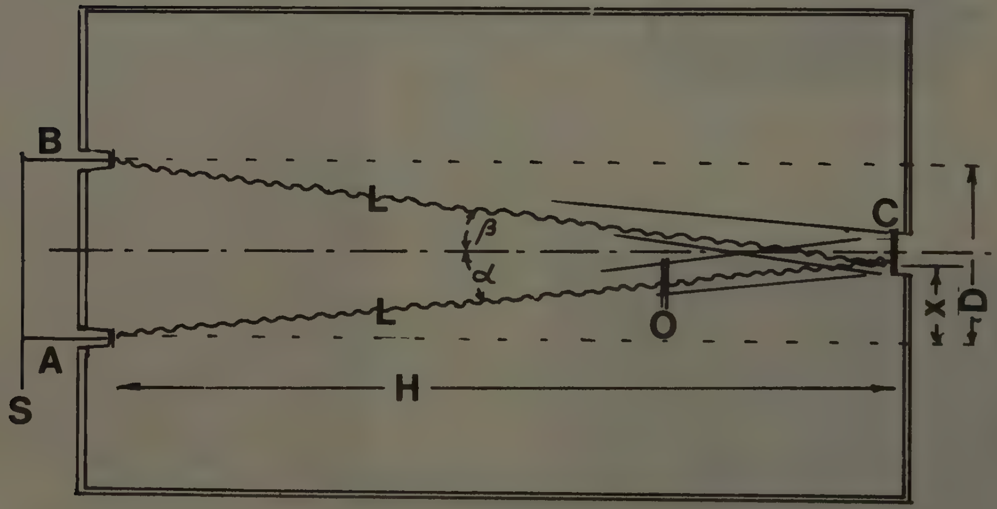


Fig. 5. The basic arrangement for ultrasonic holography with two sources, A to insonate the object O and the reference beam B. The beams subtend angles α and β to the normal to the hologram plane C. The ultrasound field is enclosed in a water bath of length H. C may be a photographic plate undergoing development which is accelerated by ultrasonic agitation; or it may be a Sokolov type of electron scanned image camera. The arrangement is not favorable because α and β are limited to about 11 degrees if shear waves are not to be developed in C. Further the object must be far from the camera if it is not to intercept the reference beam. Further neither of the transducers mentioned are efficient multiplying devices. A better arrangement is shown in Figure 6.

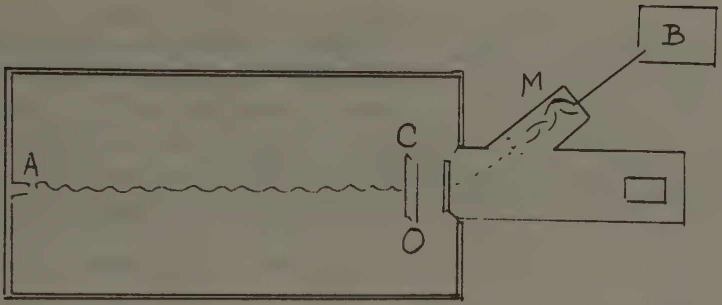


Fig. 6. Preferred arrangement for Holography

Sound source A normal to the hologram plane C. Object O close to C so that maximum scattered radiation falls on C.

Reference signal B artificially introduced into camera electron multiplier M. B and A differ in frequency by constant small amount Δw (2KHz) so that the reference beam represents a source inclined to the camera face at 60° or more. The output signal $\sin wt \times \sin (w + \Delta w)t$ contains a double frequency term which is filtered away and the Δw term which provides the total hologram information. Δw is synchronized to the frame scanning.

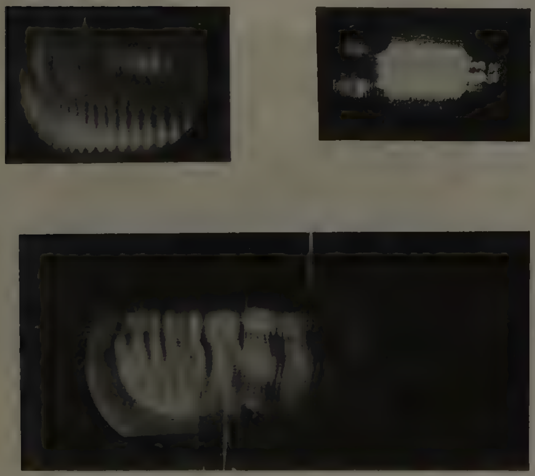


Fig. 7. Holograph distortions

Top left: Shear waves distort the parallel fringe pattern when the reference beam obliquity exceeds 11.4°.

Top right: Reconstruction of 'addition' hologram gives blurred images and central glare.

Lower: Addition hologram. 1st. and 2nd. order terms confused and fringe pattern lost.

Fourier transform of a sine wave contains only one, the second order, term. This is illustrated in Figure 8, which shows fringes obtained acoustically on the ultrasonic camera scanned at television frequencies. The star-like pictures below are the laser reconstructions of the source, obtained when the fringes are reduced photographically to a spacing of 100 lines/cm. This is a 10 to 1 linear reduction from the pattern on the camera face and leaves a 50 to 1 disparity between the optical and effective acoustic wavelengths. Accordingly, the reconstructed image is small

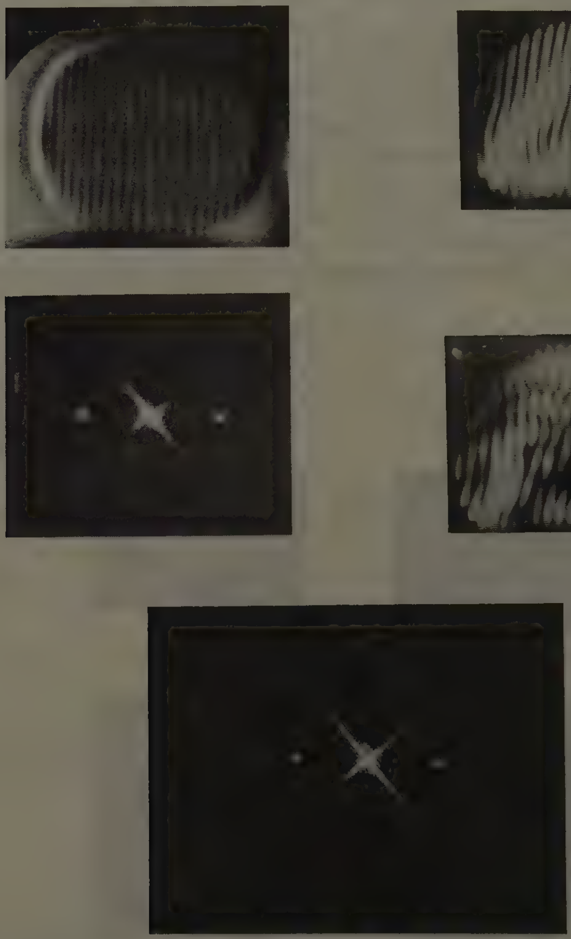


Fig. 8. Hologram parameters

Top left: Hologram 24 fringes, electronic scanning, two 4.4 MHz sources.

Middle left: Laser reconstruction of above.

Lower left: Laser reconstruction of two 10MHz sources 200 line hologram, mechanical scanning.

Top right: Sound camera hologram, wire loop.

Lower right: Sound camera hologram of acoustic lens off axis.

and appears 50 times too far away. It will be $1/50$ radian off axis. It can be viewed and photographed by a telescope or coupled to a television camera, but the distance makes the '3D', parallax, effect negligible. The sono-hologram could be further reduced optically but then imperfections in the thickness of the glass, and photographic emulsion, and variations in temperature, add imperfections to the optical reconstruction. Figure 9e and f show a cross and Fig. 10a and b show a finger as 24 line holograms and their reconstruction with a helium-neon laser. At least 200 lines are necessary to obtain pictures of the quality of Figure 9g, which was obtained by mechanical scanning.

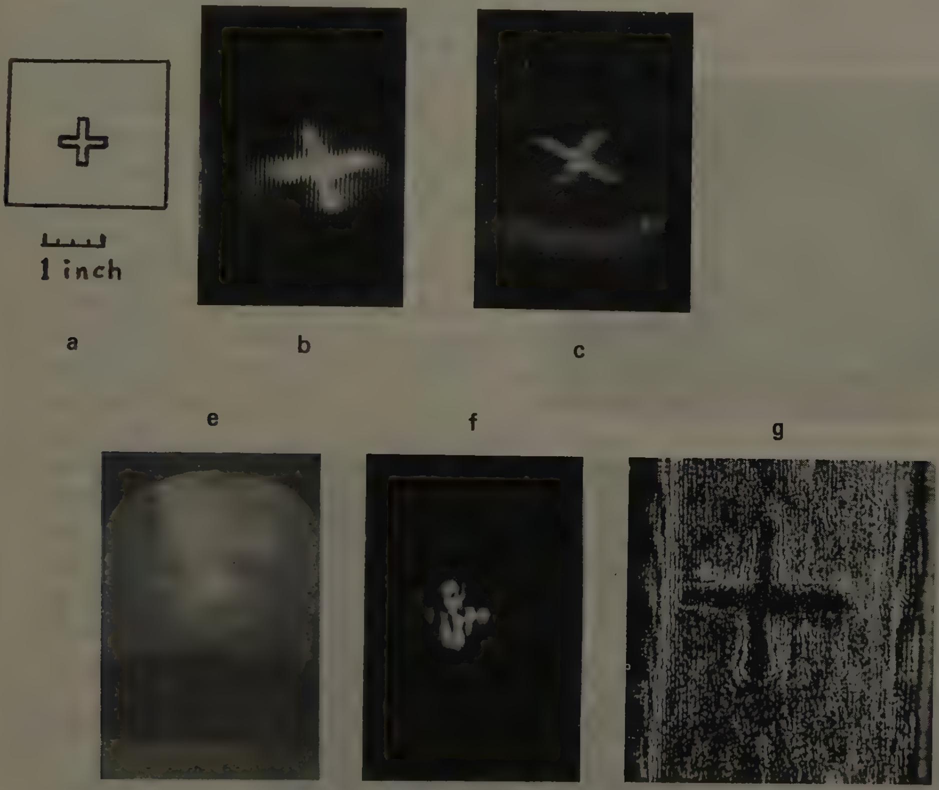


Fig. 9. Cross pattern reproduced by various techniques

- a Object
- b Acoustic shadow on image camera
- c Focused by lens on to image camera
- d phase color picture, shadow (see Color Plate p. 192)
- e Hologram, electronic scanning at 4.4 MHz
- f Reconstruction from e, blurred from poor CRT scan linearity
- g Reconstruction from e, blurred from poor CRT scan linearity (multiplier circuit) 200 line hologram.

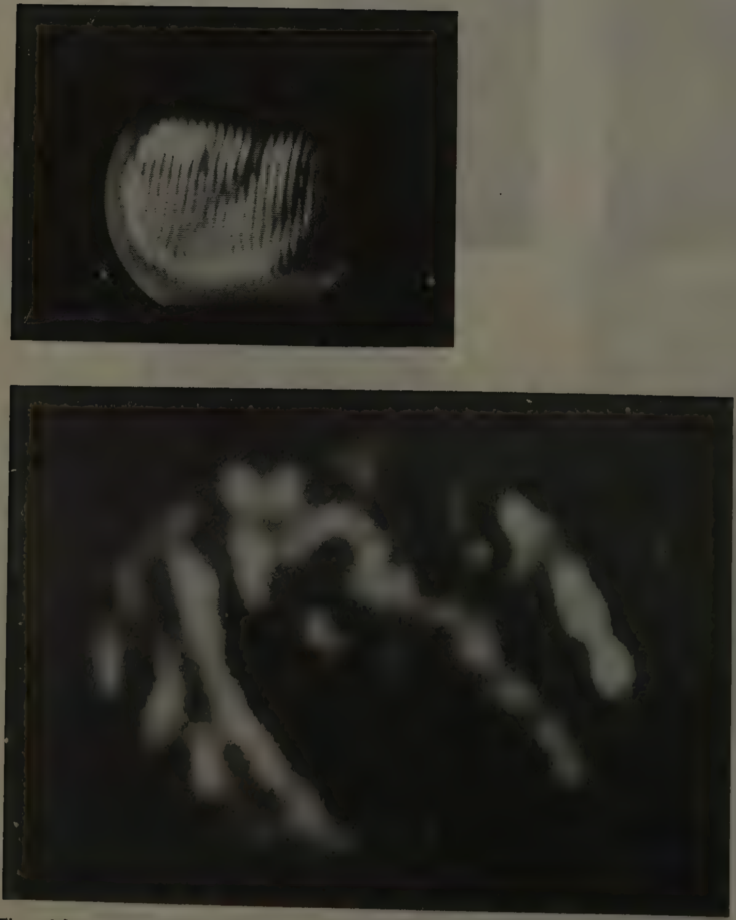


Fig. 10.

a. Sound camera hologram of finger (4 MHz, 24 line).

b. Reconstruction.

(see Color Plate p. 192 for Fig. 10.c)

The use of an electrical reference signal to provide the phase information is achieved by applying an alternating potential to the electron multiplier or to one section of the multiplier chain. The output contains the product of input signal and the reference wave; there is a double frequency term and the power or phase term. The amplifiers and viewing tube disregard the high frequency component. The phase pattern, if the picture is suitably biased, varies the brightness from black to white at the appropriate picture points.

If the reference signal is the same frequency as the object insonation, then the reference signal is orthogonal to the holograph plane. If the phase of the reference signal varies continuously with the frame or line scanning, then there is an effective obliquity of the reference signal. The simplest method to achieve this is to have the reference signal at a different frequency from the source signal such that the number of cycles per scan difference is equal to the number of fringes across the picture. If 40 lines are required in 1/50th second the frequency difference must be accurately maintained at 2 KHz. The frame scanning can, with advantage, be synchronized to this difference frequency.

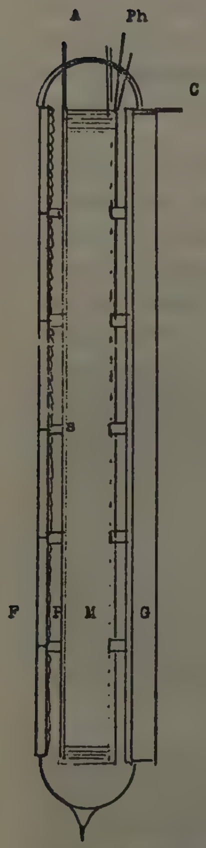
It will be appreciated that the linearity of scanning must be of a high order if the fringes are to be equally spaced, for if not the reproduced picture will be blurred (Fig. 9f).

With the electronically operated image camera, the object signal must arrive at the quartz plate at an incidence of less than 11.4° if the production of shear waves is to be avoided. It is therefore advantageous to use this cone maximally and arrange the main axis of the object insonation normal to the camera face. The reference beam will be at perhaps 60° to the normal, and set the hologram limiting resolution above that possible from the object (22.8°). With artificially produced reference beams the fringe spacing can be reduced to the finest limits reproducable on the viewing screen; the implications of this for picture quality have not yet been explored. Figure 6 is an example of such a hologram arrangement. Figure 7 shows the unacceptable shear waves. With holograms, as with direct images, the signal can be pulsed and gated to produce tomogram sections at the cost of slowing down the process. To regain the speed, direct image methods are necessary.

Holography by Direct Image Methods

The method described here is as yet not constructed; it is illustrated in principle in Figure 12a. The design is based on the Mullard channel multiplier matrix (14). For an oblique signal and a normal reference beam, the tube is simple to construct, but it would be better to provide an oblique reference and a normal signal beam. To produce an oblique reference beam the multiplier plate must be metalized in parallel stripes at least four to a phase wave and these sequentially connected to a delay line or phase splitting transformer. With such a system, the whole hologram could be made in less than four cycles of the ultrasonic frequency, say a microsecond and body movements would not be consequential.

Optical image intensifiers have been made on this principle and Figure 12b shows the resolution obtainable.



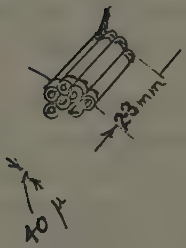


Fig. 12a. Design for channel amplifier tube

Section of ultrasonic image converter based on design of commercial optical image converter using Mullard channel electron multiplier matrix.

Above: detail matrix.

Symbols

F	Quartz sectional camera face plate
P	Photomosaic, gold, on F
M	Channel matrix
A	Accelerator electrode and channel input end electrode metallised on matrix and photoconducting to provide electrons for recharging the mosaic.
Ph	Phase electrodes for holography
S	Structural supporting pillars
G	Phosphor coated conducting glass
C	Output electrode.

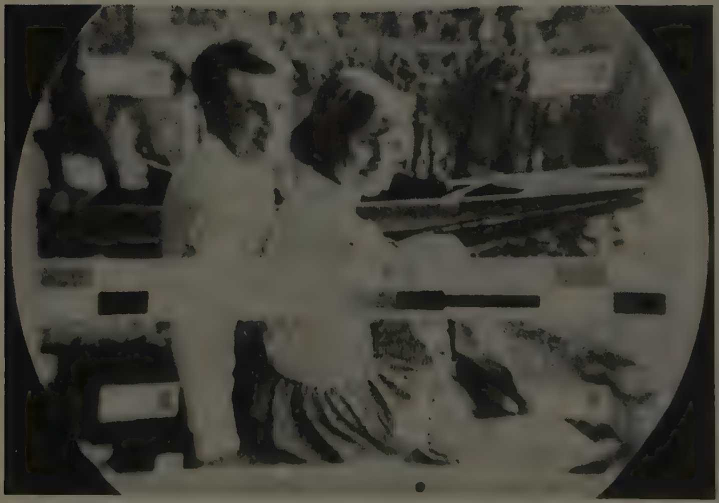


Fig. 12.b. Optical resolution obtainable with channel multiplier.

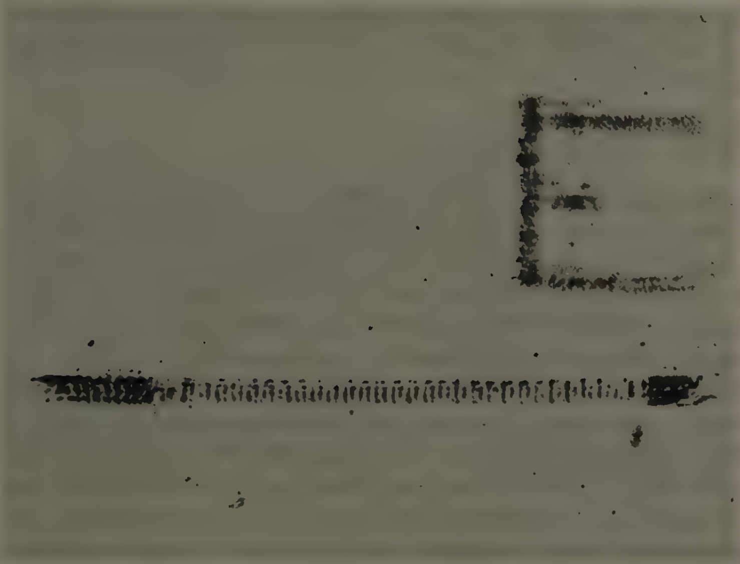
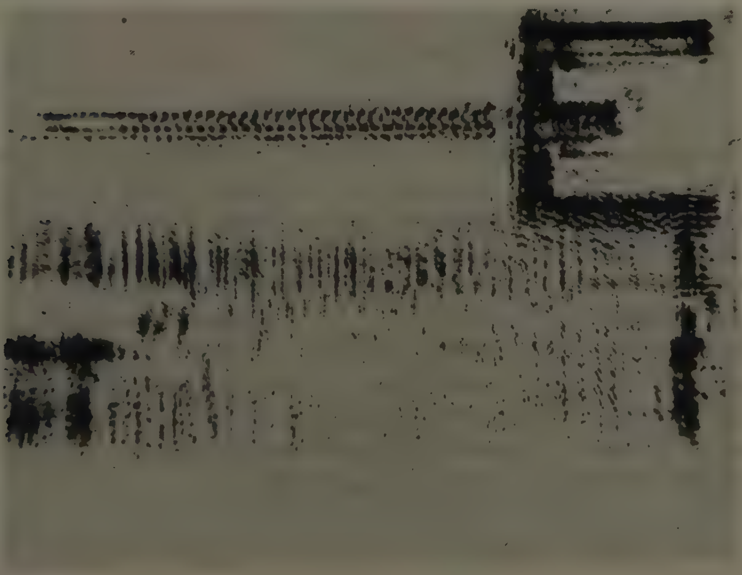


Fig. 13. 10 MHz mechanically scanned hologram. Reconstructions illustrating change of focus by 3 cm.

a) focus on the screw.

b) focus on the 'E'.



One advantage of this structure is that the channel multiplier block would provide support for a much larger quartz plate and the picture size could be increased. Further, with holograms, several plates could be used to reduce shear wave effects and the gaps between would not appear in the reproduced pictures. To increase the angle and resolution, the surface of the quartz elements might be lenticulated.

Holography by Mechanical Scanning*

The apparatus comprises a mechanically scanned system in which a plane isonifying beam is synthesized by scanning a focused transducer across the aperture, and a plane inclined reference beam is synthesized electronically and combined with ultrasound scattered by the object in a linear multiplier. The hologram is recorded on a facsimile recorder and the optical transparency is made from this. Figure 9g shows the type of image quality being produced at 10 MHz.

A hologram was made of a letter 'E' in a 1 cm square with arms 1½ and 2½mm wide in brass sheet 2½mm thick held in position with 6BA studding at 4cm from the holograph plane. In front of this and 1cm from the holograph plane was a second object, a 2BA nylon cheese head screw with the head laterally level with the 'E' and again supported by 6BA studding. The ultrasound was pulsed on for 10 microseconds with a repetition rate of 700 Hz. The reproductions show the screw in focus, Figure 13a, while Figure 13b is focused on the 'E'.

---

* Communication from E. E. Aldridge, A.E.R.E., Harwell.

ACKNOWLEDGMENTS

We are grateful to Professor F. Y. Poynton for facilities at the City University where research has been conducted on this subject since 1954, and for his continued encouragement and assistance with theoretical problems. Past work there was supported by the R. W. Paul Instrument Fund and recent work there has been supported by the Atomic Energy Research Establishment, Harwell. Mr. E. E. Aldridge of that establishment has advised us on holographic procedures and kindly allowed the pictures of some of his work to be reproduced in this article, Figs. 8 (lower right), 9g, and Fig. 13. With Mr. A. Clare he made laser reconstructions of our sound camera holograms. We are indebted to the editor of The Journal of the Institute of Electrical Engineers for permission to reproduce figure 1b, and to the editor of the Mullard Technical Bulletin for Fig. 12b.

REFERENCES

This list has been restricted to relevant original papers or articles that provide further extensive source information.

1. Lipson, H., Taylor, C. A. Fourier Transforms and X-ray- Diffraction, pub. Bell, London, 1958.
2. Gordon, D. Ultrasonics as a Diagnostic and Surgical Tool, pub. Livingstone, Edinburgh, 1964.
3. Smyth, C. N., Poynton, F. Y., Sayers, J. F. The Ultrasound Image Camera. *Proc. I.E.E.* 1963, **110**, pp 16-28.
4. Gabor, D. *Nature* **161,** 777 1948 and ref. 12, Ch. 18., 1969.
5. Sokolov, S. Ya. 1936 USSR Patent No. 49 of Aug. 31, 1936; see also references in (3).
6. von Fritag, W., Martin, H. J. Jena Jarhbuch 1958 p. 228.
7. Jacobs, J. E. 1968 I.E.E.E. Trans. Ultrasonics Eng. UE 14.
8. Smyth, C. N. vide 3 above p. 23, Fig. 11.
9. Jacobs, J. E. *Materials Evaluation,* August 1968. p. 158,
10. Fritzler, D., Marom, E., Mueller, R. K., Acoustic Holography, 1969, pub. Plenum Press, Ch. 16.
11. Aldredge, E. E. *Physics Bulletin,* Vol. 20, Jan. 1969, pp 10-12.
12. Acoustic Holography Conference Report, 1969, Plenum Press.
13. Smyth, C. N., J.I.E.E. Vol. 9 Jan. 1963 paper 4030E.
14. Guest, A. Holmshaw, R. I. Manley, B. W. Channel Multiplier Plates for Imaging Application. *Mullard Technical Communications.* T.P. 1064, 1968, Mullard, Ltd., London, W.C.I.

ULTRASONIC IMAGING AND DECOMPRESSION SICKNESS STUDIES

R. Stuart Mackay

A potential problem faced by divers in the ocean, airplane pilots, tunnel workers, and practitioners of hyperbaric medicine is that of decompression sickness or the "bends". In this, too rapid a transition to a region of lower atmospheric pressure can cause the formation of bubbles by the evolution of the inert gases in the breathing mixture from solution in the tissues and blood. The result can be anything from an itching sensation through pain and death. The breathing of an oxygenated liquid potentially removes this problem, but it will be some time before this can be considered practical even for the most rigorous procedures. The present review will give some of the details of the problem and the contributions that can be made in basic studies, as well as in preparing diving tables and guiding the decompression of actual individuals, by the ultrasonic detection or visualization of bubbles that may form in the body.

The hazards of rapid decompression have been known to divers for many decades. Haldane proposed a model to describe these matters in a quantitative way. It is perhaps one of the oldest biological models actually used in the conduct of human affairs. In essence, it describes the body as made up of a collection of regions or compartments that dissolve or evolve the inert gas in a breathing mixture with characteristic half-times of 5 minutes, 10 minutes, 20 minutes, 40 minutes, 80 minutes, 120 minutes, and perhaps more. The rate of change of the partial pressure of the gas in a given tissue is taken as proportional to the difference in the partial pressure of that gas in the breathing mixture and in the tissue. It is further assumed that a bubble will form when the partial pressure π in a given tissue becomes roughly twice (γ, depending on the tissue type) the pressure of the breathing gas (which must equal the ambient pressure P). It should be noted that one cannot do away with the inert gas because at elevated pressures, pure oxygen acts as a poison. One can improve the situation slightly by breathing pure oxygen at atmospheric pressure before starting an activity in order that the various tissues will start with no dissolved inert gas.

$$\frac{d\Pi_i}{dt} = K_i\,[P_N - \Pi_i] \qquad \text{where } K_i = .693/T_i \qquad \text{and } T_i = 5, 10, 20, \text{etc.}$$

$$\text{for a bubble } \Pi_i = \gamma_i P \qquad \text{where } \gamma_i \sim 2$$

For optimum use of a given inert gas, use a readily obtained continuously variable gas mixture containing as much oxygen as allowable, in some cases two atmospheres: $P_N = P - 66$ where pressures are in feet of water.

Solutions to the equations involved have been given, along with some results generated with a simplified analog computer arrangement (1). This computer arrangement has proven valuable in understanding and visualizing the effect of a change in diving conditions such as depth, or time, or gas mixture, or repeated dives, etc. One connection is reproduced in Fig. 1 for the benefit of those who might wish to use it as an instructional aid in a diving program, or who might better be able to understand the model from these interconnections.

This connection simulates a diver going to a depth of 200 feet (a pressure of 233 feet absolute). Values taken for the actual allowable overpressure in the different tissue types as calculated back from a diving table (1) are: 2.7, 2.3, 2.0, 1.7, 1.6, and 1.6. The diver starts his descent breathing pure oxygen, and at all other depths, breathes a mixture of 2 atmospheres oxygen and the rest nitrogen to make up a mixture equal in pressure to the ambient pressure. With descent at 125 ft. per minute, 20 minutes on the bottom, and initial ascent at 100 ft. per minute,

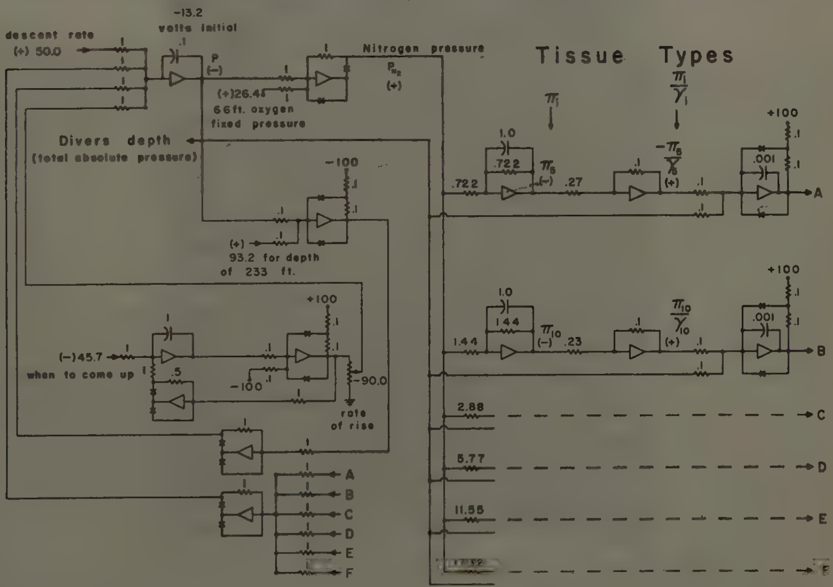

-13.2
volts initial
descent rate
(+) 50.0
P
(–)
Nitrogen pressure
P_{N_2}
(+)
(+)26.4
66 ft. oxygen
fixed pressure
Divers depth
(total absolute pressure)
-100
(+)
93.2 for depth
of 233 ft.
+100
(–)45.7
when to come up
-100
-90.0
rate
of rise
A
B
C
D
E
F
Tissue Types
π_1
$\frac{\pi_1}{\gamma_1}$
1.0
.722
.722
π_5
(–)
.27
$-\frac{\pi_5}{\gamma_5}$
(+)
+100
.001
A
1.0
1.44
1.44
π_{10}
(–)
.23
$\frac{\pi_{10}}{\gamma_{10}}$
(+)
+100
.001
B
2.88
C
5.77
D
11.55
E
F

Fig. 1. Analog computer interconnections for calculating and visualizing the decompression required by a diver or other subject going to a region of reduced pressure. Similar models can apply to anesthetics, the simulation of kidney function, radioactive tracers, etc.

the indicated total time for return to surface is approximately 9.7 minutes, as compared with a required time of 40 minutes as indicated by the standard stepwise decompression table (from which the allowable overpressures were computed) for air breathing. Computations can be made in real time, but in this configuration, one second represents 10 minutes and 0.4 volt represents 1 foot. To compute for air breathing rather than the continuously variable gas mixture, the upper center configuration is merely replaced by a 4/5 voltage divider to indicate that the nitrogen pressure is this fraction of the total pressure, which is itself the diver's depth.

The curves produced by such configurations are indicated in Fig. 2, in which the diver's position is indicated by pressures increasing upward. Superimposed are the curves Π/γ for three of the tissue types. The pressure on the diver rises to beyond the present figure and then remains constant for the useful period on the bottom. The pressure then drops linearly as he returns toward the surface, and this portion of the ascent is interrupted by a tendency for the most rapid tissue type to bubble. Successive tissue types successively limit ascent, and it is seen that the diver must successively follow the successive Π/γ curves until the surface is reached. The surface is indicated by a pressure of 33 ft. of water (one atmosphere), and restrictions on then taking a helicopter ride can be followed from the remainder of the trace. In viewing such a display, it becomes quite clear how a change in diving conditions will alter the required decompression stops in a diving table. For further information, the reader is referred to the original article.

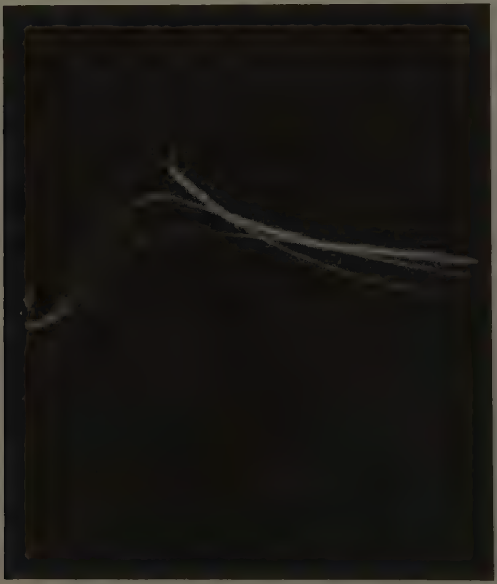


Fig. 2. Oscillographic recording on an enlarged scale showing the depth of a diver as increasing pressure upward, as a function of time. Superimposed are the curves for three successive tissue types showing the changes in dissolved inert gas. During the interesting part of the ascent, the diver must follow the curves in succession as each verges on bubbling. Note that the diver starts from a pressure of 33 ft. (one atmosphere), while the initial tissue pressure of nitrogen is zero.

Diving experiments can be considered as determining allowable values of γ. In studying decompression sickness or preparing a diving table, one has had to use human subjects who can be asked if they hurt following some decompression schedule. The onset of pain may not be immediate, and it has never really been demonstrated that bubbles are present at the actual site of pain. In any case, a more objective indication of bubbles would be desirable, and preferably one that could also be applied to animal subjects, perhaps free-swimming. With the perfection of telemetry techniques for communicating information from unrestrained subjects (2), this could become especially desirable. These techniques also allow monitoring without interfering with the circadian rhythm in such parameters as blood flow rate, which indeed might affect decompression schedules following saturation dives. The remainder of the discussion will be devoted to a method for implementing such studies and observations.

It has become possible to build up images of internal body structures, especially in the soft tissues, using pulsed ultrasonic energy in a sort of high frequency scanning sonar system. There need be no damaging effects from repeated observations, such as might accompany a similar observation using X-rays. As an example of the detail that can be seen even in a densely packed structure such as the human neck, Fig. 3 reproduces an image supplied by one of the pioneers in this field, the late Douglas Howry. The echoes returning to the scanning transducer are reflections from the interfaces between differing tissue types. An interface between

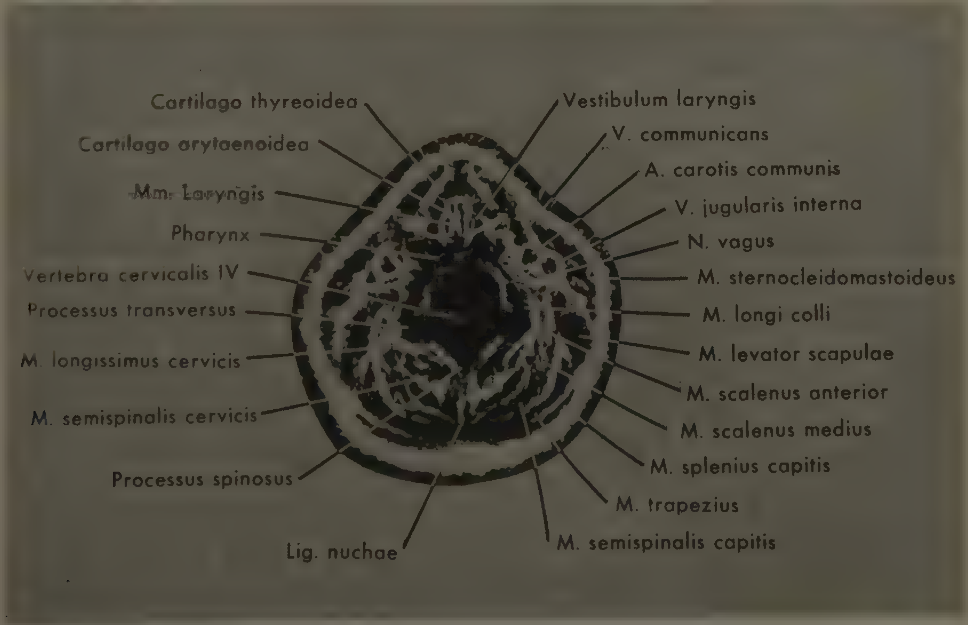

Cartilago thyreoidea
Cartilago arytaenoidea
Mm. Laryngis
Pharynx
Vertebra cervicalis IV
Processus transversus
M. longissimus cervicis
M. semispinalis cervicis
Processus spinosus
Lig. nuchae
Vestibulum laryngis
V. communicans
A. carotis communis
V. jugularis interna
N. vagus
M. sternocleidomastoideus
M. longi colli
M. levator scapulae
M. scalenus anterior
M. scalenus medius
M. splenius capitis
M. trapezius
M. semispinalis capitis

Fig. 3. Ultrasonic image of the densely packed structures in the human neck, as produced by Douglas Howry.

a gas and a solid or liquid is an especially good reflector because of the great difference in acoustic impedance (product of density and sound velocity). Fig. 4 reproduces the well-known reflection properties of small objects for sound waves. At a frequency of 15 megacycles, the resonant size of a bubble is a little under 1/2 micron.

Accordingly, it seemed reasonable to ascertain if a sound signal could detect bubbles produced by rapid decompression. At the Second Symposium on Underwater Physiology in 1963, it was noted (3): "Much of the discussion has emphasized a fundamental need in supersaturation and bubble formation experiments for a definite indication of endpoint, and there may be a new approach. It is the use of ultrasound to detect the presence of tiny bubbles and the onset of the bends. In our laboratory at Berkeley, we have built pulsed 15-megacycle sound echo-exploring equipment, and gas bubbles are excellent reflectors. At this frequency, the wavelength is 100 microns, and these circuits can detect and follow the progress of single bubbles that are an appreciable fraction of this dimension, or larger. In the preliminary experiment, I demonstrated the change in properties (increased opacity) to ultrasound that rat tissue undergoes upon being given vigorous decompression sickness. A focused transducer and simple scan were used, but a bulk obser-

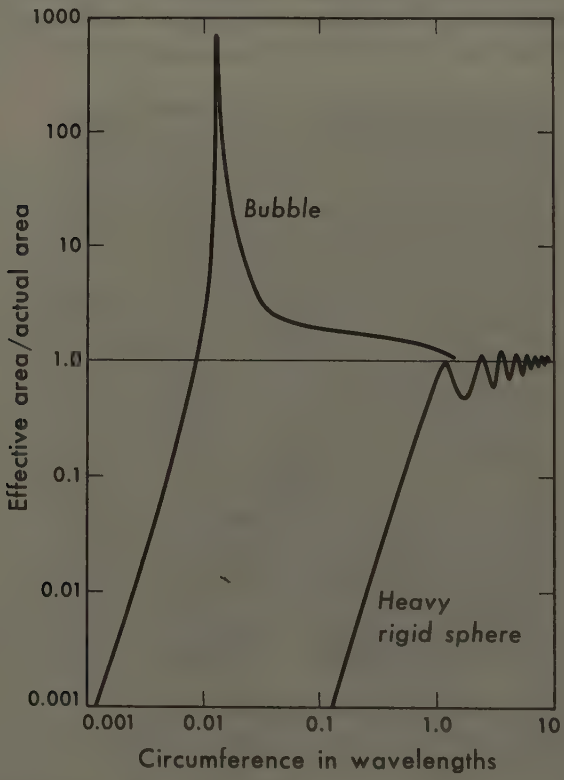


Fig. 4. Scattering of sound waves by small objects depends upon the relative size of the object and a wavelength.

vation with continuous sound also should have been effective. Either sound transmission or back-scattering methods could be employed, and they would seem to provide an excellent tool both for *in vivo* and *in vitro* studies of the kinetics of bubble formation, for example with mixed inert gasses. Some workers speak of 'silent bubbles' and these should be detectable, if real. It might be mentioned that the sound intensities here are very low (about 0.001 watt/cm^2). Rapid excess inward diffusion of gas on a rarefaction half-cycle might allow intense sound to 'pump up' otherwise unnoticeable small bubbles, in a fixed way, until detectable. I would imagine that a diver who was decompressing, and was hit by an intense sonar beam, might be given a case of the bends. We could investigate the effect of intense ultrasonic energy on a tissue having an overpressure with our other therapeutic equipment, but I have not yet done this; the degree of supersaturation might thus be measured."

The electronics of such an imaging system need not be complex, though there are some problems associated with the fact that the electrical signal of several hundred volts that generates the outgoing sound pulse is also applied to the input of the sensitive amplifier that will later receive the small echoes returning to the same transducer. The mechanical system must be quite sturdy if the expected resolution is to be realized during scanning. Figs 5 and 6 (reproduced from Ref. 4) indicate some details of the scheme and one of its components. Fig. 6 shows the focusing action of a biconcave lens for sound energy, as visualized by placing the sound source in water and directing it across a Schlierin light-optical system. More will be said of this method of converting a sound image to a light image later.

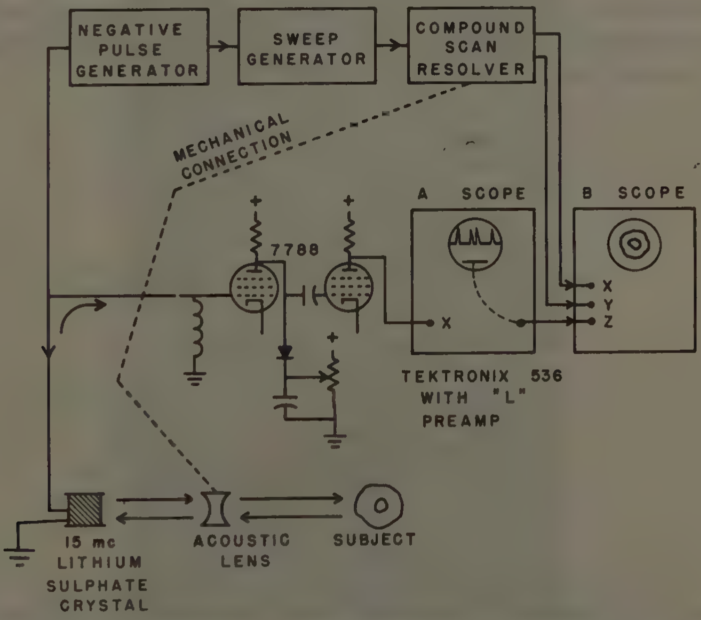


Fig. 5. Pulsed ultrasonic scanning system using standard oscilloscope for the displays.

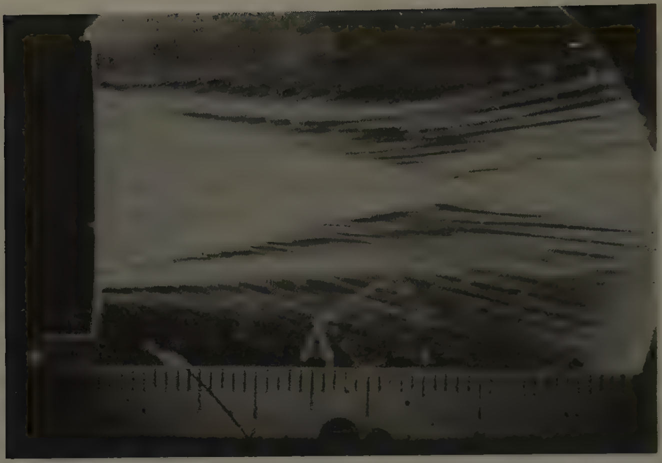


Fig. 6. Focusing action of an ultrasonic lens, the sound pattern being converted into a visible image by a Schlierin optical system.

The use of phase information in the returning echoes might also be used to selectively emphasize regions of free gas (Fig. 7). On a color display, for example, any large amplitude echo of reversed phase from that of the outgoing signal might be displayed in red, while echoes of the original phase could be in blue. In the display oscilloscope, not intensity changes but small vertical deflections can be applied to the scan pattern to give an appearance that is different in a useful way; up and down then distinguish the different returning phases of echoes. Similarly, a continuous outgoing signal mixed with some of its second harmonic will undergo a change in wave shape after reflection from a more-to-less dense interface, and this could also be used to sense extended gas regions.

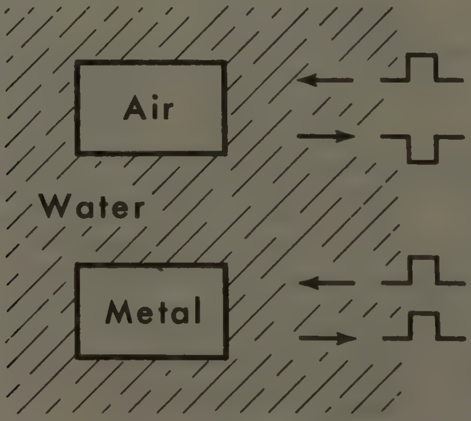


Fig. 7. Phase information can be used to distinguish reflections at different types of interfaces. Depicted here is the inversion of a pulse, with the returned pulse perhaps having the same frequency components in both cases.

The above was discussed in a number of publications, but interest mounted slowly until recently (5-12). The piezoelectric transducers can readily withstand rather large pressure changes. However, in work with George Rubissow, it was decided to place the scanning system outside the decompression chamber, and to form images through a thin Mylar window that would withstand the pressure changes while not introducing distortion into the sound pattern (12). The apparatus and some representative results are shown in the next four figures (Figs. 8 through 11). In Fig. 8 is seen the chamber with the acoustic window in the side, and a trough of water for good acoustic coupling in which the transducer scans twenty times per second. The unit pulses 3000 times per second, and is supported on a modified drill press, as in Fig. 9. Observations on the leg of a small mammal are depicted in Figs. 10 and 11.

Workers in England (5) noted an increase in opacity after a time delay, that was reversible upon recompression, of a reflection from the back surface of a fold of tissue while using a commercially available pulsed ultrasonic system that did not scan. Other workers there have suggested looking for an increase in returned second harmonic in the presence of bubbles (6). The suggestion is of interest, though there may be problems with sufficient second harmonic generation at low sound intensity, and nonlinearities in the rest of the overall system might also contribute some confusing second harmonic.

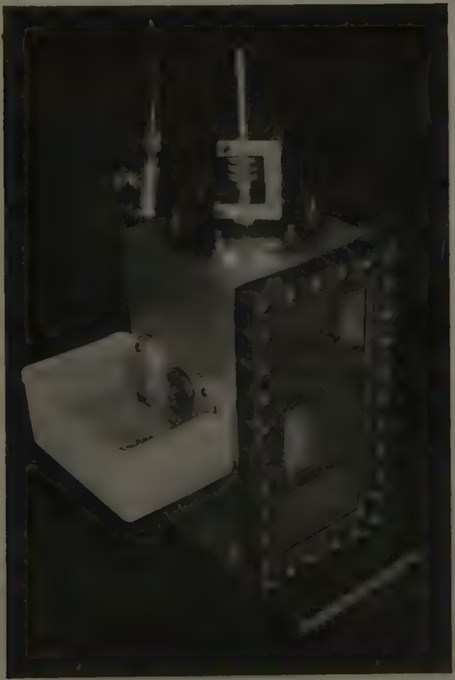


Fig. 8. Pressure chamber with acoustic window in left side. At top is the rod for manipulating the animal in the chamber.

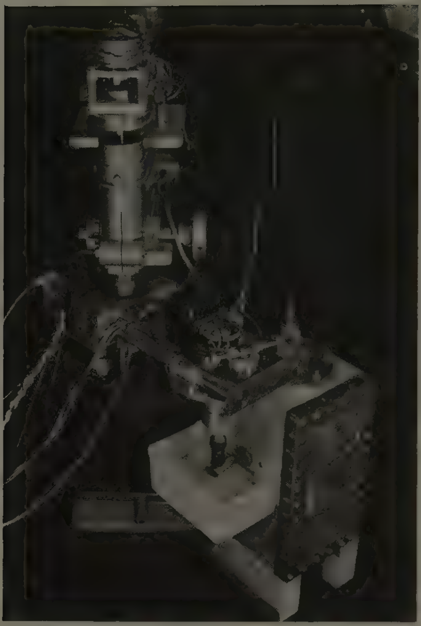
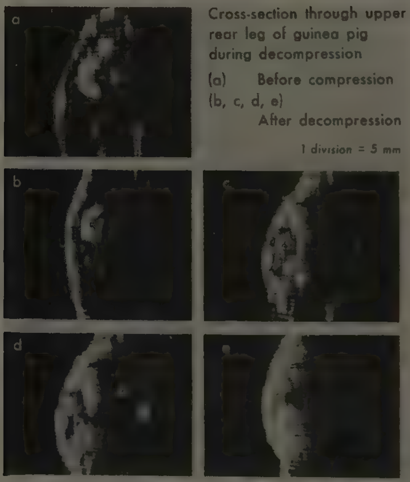


Fig. 9. Ultrasonic scanning equipment with transducer in trough of outgassed saline at the side of the pressure chamber.



a
Cross-section through upper rear leg of guinea pig during decompression
(a) Before compression
(b, c, d, e) After decompression
1 division = 5 mm
b
c
d

Fig. 10. Image through upper leg of guinea pig, with sound source to left. Before decompression, the flattened bright area of bone can be seen, as well as the various surfaces. After decompression, the remote surfaces become obscure, though at *d*, bubbles can be seen forming on a remote surface opposite a region where are fewer bubbles internally. The last four images are at 1 minute, 1 minute 30 seconds, 1 minute 35 seconds, and 1 minute 50 seconds, after abrupt return to atmospheric pressure.

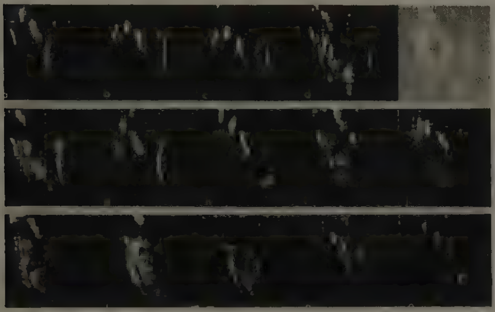


Fig. 11. A simple scan was used to emphasize bubbles to the exclusion of body structures during decompression and subsequent therapeutic recompression of a guinea pig. The crescent is the doubly reflected image of the sound window bounced off the transducer. Upon recompression, there was an immediate decrease in the bubbles followed by further gradual decrease. During the final decompression back to the original condition, the appearance of the bubbles varied with each pressure change.

A blood flowmeter was demonstrated (13) and later became commercially available, which operated by observing the Doppler shift in frequency in sound scattered backwards from the formed elements (red cells) in blood. The scheme is indicated in Fig. 12, where a 10 megacycle sound signal is directed into a blood vessel. (Operation through the intact skin is possible with a change in geometry.) At the lower transducer, the scattered sound at a slightly different frequency will combine with some direct sound at the original 10 megacycle frequency to give a beat frequency proportional to flow velocity. This can be detected and telemetered as indicated, or handled by various other frequency analyzing circuits.

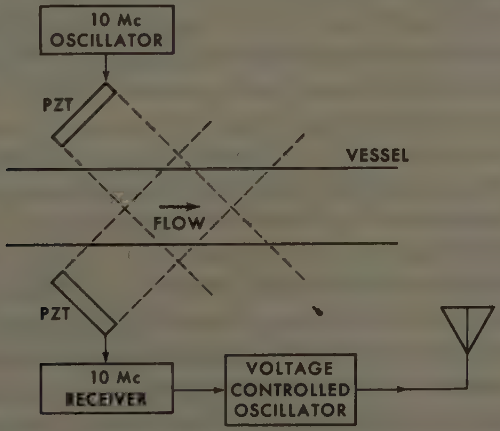


Fig. 12. Scheme of Doppler shift flowmeter. Though normally used to measure the velocity of blood flow, these are very sensitive to the presence of moving bubbles.

When such a device is applied to an animal undergoing rapid changes in pressure, there can be a series of annoying clicks. This fact was observed and exploited by several workers. Gillis et al. (7) used a Doppler unit to observe circulating emboli in swine vena cava before the first decompression stop recommended for man. There was occasional total signal loss. Transcutaneous observations were also made by them on dog, swine, and a goat (8). Spencer and Campbell (9) observed "whistles" and "roars" from the inferior vena cava and the descending aorta of sheep during decompression. They also found "artifact voltages" from an implanted electromagnetic flowmeter simultaneous with the Doppler-detected bubbles. They found bubbles in the venous blood of sheep at pressures greater than the usual first decompression stop for man.

The extent to which sounds in the 10 megacycle frequency range expedite bubble formation is not clear, but if bubbles only are to be detected, then sound intensities can be reduced. The Doppler systems display information about the existence of moving structures only, and though this reduces the receipt of confusing echoes, it makes it impossible to follow a given bubble or to tell where a bubble is, and if a bubble expands to fill a part of the vascular system or becomes stationary, it then becomes invisible. Similarly, possible bubbles in tissue cannot be studied. Pulsed Doppler systems can be built to give position and approximate velocity, but they are generally not needed in this application.

The pulsed ultrasonic systems give a useful cross-section image by their scanning action. There are other procedures for converting an ultrasonic image into a visible one, and some can employ pulsed energy to visualize structure at a given range. For example, Smyth, Jacobs, and others have built quartz crystal-fronted television tubes that would convert a sound pattern into a video image. Such devices might also be used to record holograms from the south field. These tubes accommodate rapidly changing images, but to date, have been of small extent and limited resolution (14).

Fig. 6 also depicts the conversion of a sound pattern into a visible one. In this case, the sound pattern existed in a glass-sided tank of water. Light shining across the direction of sound propagation made the pattern of density fluctuations visible through the action of a Schlierin optical system. An increase in light intensity would allow the sound intensity to be reduced. The brightness of even a low powered laser can be extremely high, and thus the use of one of these as the light source in such a system should make it possible to study biological subjects in this way. It has also been shown that in an experiment of similar gross geometry, with a converging monochromatic light beam, Bragg defraction occurs and a reduced optical image of the original object is formed (15). This has recently been used to form images of the leg of a small mammal during decompression (11).

Sample bubbles for study can be produced by electrolysis, by droplets of peroxide, or by the turbulent injection of one liquid through a tube into another.

A familiar observation by radiologists is the sudden change in appearance of the joint of a human during movement. Thus moving the legs of an infant can cause the sudden brightening of the hip joint on a fluoroscope screen. Moving the arms sometimes causes a similar change in the shoulder, as does "popping" of the knuckles. Such methods as the above will demonstrate if this is truly the release of a bubble of gas during normal activity. If so, it could have implications for the often reported "joint pain" of divers.

With these methods, even the eye can safely be studied. Thus similar methods should allow evaluation of changes in the elastic properties of the lens (2) for assessing long-term disturbance by breathing excess oxygen, etc.

Most ultrasonic equipment can function effectively inside decompression chambers, if this is desirable. However, in studying human subjects for bubble formation, it may be sufficient to extract blood which has equilibrated with tissues from the vena cava through a cuvette heated to 37° C. In diving experiments to determine allowable overpressures in different tissues, this criterion of bubble formation might be simpler and less subject to statistical variation than more subjective tests. An uncertain part of a diving table might selectively be checked by a few precise tests. Too fast a decompression at late stops presumably would hit regions of low perfusion in the body. Inducing bubbles in a known way with more intense ultrasound might allow measurement of the degree of supersaturation in a given region. It should be possible to tell if the ratio of pressures or the difference in pressures is actually the most significant, and generally to examine the fundamental postulates of the onset of decompression sickness. It should be possible to further explore the existence of "silent bubbles", and to tell what size a bubble must be to exercise an effect. It seems possible to have bubbles without pain, and it may be that there can be pain without bubbles at the actual site.

It may be that any bubble which forms will expand dangerously upon further decompression. Thus such methods may prove useful in guiding the ascent of actual divers by modifying the subsequent decompression schedule following bubble onset to match the momentary overall condition of a given diver. From long-term saturation dives, it may be possible to tell if altered blood flow patterns while asleep and awake during decompression demand periodically changed schedules. Successes in transmitting information through a watery medium (2) suggest that this objective test for bubbles may be applied to free swimming animals and humans in future experiments, and probably will settle the classic questions about why whales do not suffer from decompression sickness by indicating if they may not under some circumstances.

REFERENCES

1. Bradner, H. and Mackay, R. S., Biophysical limitations on deep diving: Some limiting performance expectations. *Bulletin Mathematical Biophysics* **25,** pp. 251-272, Sept. 1963.
2. Mackay, R. S., *Bio-Medical Telemetry: Sensing and Transmitting Biological Information from Animals and Man.* John Wiley & Sons, Inc., New York, 1968.
3. Mackay, R. S., In *Proc. Second Symposium on Underwater Physiology.* (Eds: C. Lambertsen, L. Greenbaum) Washington, D. C. National Academy of Sciences/National Research Council Publication 1181, 1963, p. 41.
4. Mackay, R. S., Telemetering physiological information from within cetaceans, and the applicability of ultrasound to understanding *in vivo* structure and performance. In *Whales, Dolphins, and Porpoises.* (Ed: K. S. Norris) University of Calif. Press, Berkeley, 1966, pp. 445-470.
5. Walder, D. N., Evans, A., Hempleman, H. V., Ultrasonic monitoring of decompression. *Lancet,* Apr. **27,** 1968, pp. 897–898.
6. Tucker, D. G., Welsby, V. G., Ultrasonic monitoring of decompression. *Lancet,* June **8,** 1968, p. 1253.
7. Gillis, M. F., Karagianes, M. T., Peterson, P. L., Bends: Detection of circulating gas emboli with external sensor. *Science* **161,** pp. 579-580, 1968.
8. Gillis, M. F., Peterson, P. L., Karagianes, M. T., *In vivo* detection of circulating gas emboli associated with decompression sickness using the Doppler flowmeter. *Nature* **217,** pp. 965-967, 1968.
9. Spencer, M. P., Campbell, S. D., Development of bubbles in venous and arterial blood during hyperbaric decompression. *Bulletin Mason Clinic* **22,** pp. 26-32, 1968.
10. Kidd, D. J., Elliott, D. H., Clinical manifestations and treatment of decompression sickness in divers. In *The Physiology and Medicine of Diving and Compressed Air Work.* Baillière, Tindall and Cassell, Ltd., London, 1969.
11. Buckles, R. G., Knox, C., *In vivo* bubble detection by acoustic-optical imaging techniques. *Nature* **222,** May 24, 1969, pp. 771-772.
12. Mackay, R. S., Rubissow, G., Detection of bubbles in tissues and blood. In *The 4th Symposium On Underwater Physiology,* Academic Press, New York, (Ed: C. Lambertsen) 1970.
13. Franklin, D. L., Schlegel, W., Rushmer, R. F., Blood flow measured by Doppler frequency shift of back scattered ultrasound. *Science* **134,** pp. 564-565, 1961.
14. Mackay, R. S., Editor. *Ultrasonic Imaging.* Washington, D. C. National Academy of Sciences/National Research Council Publication NRC:MAC:2016, 1965.
15. Karpel, A., *Applied Physics Letters* **9,** p. 425, 1966.

PHOTOCHEMISTRY OF EXCITED STATES OF DNA COMPONENTS

H. E. Johns

INTRODUCTION

It has long been known that electromagnetic radiation perturbs biological systems. This fact is exploited by the radiotherapists in treating tumours with ionizing radiation, since it appears that radiations damage tumour cells slightly more than normal cells. In spite of the fact that many workers have studied the biological effects of radiation, we still do not understand the details of the action of ionizing radiation on cells or parts of cells. One of the main difficulties is the complex nature of the initial radiation damage. X-ray photons have enough energy to break any bond in any molecule in a cellular system, so that the damage is very nonspecific and many types of primary damage are produced. In recent years, some advances have been made in this field through the technique of pulse radiolysis where a pulse of ionizing radiation is put into the system, usually with a linac, and the immediate radiation products studied as they attack the system (1).

Ultraviolet light has also been used to perturb biological systems (2, 3, 4, 5). The action of UV photons is somewhat different from x-ray photons since the former are absorbed in specific bonds in a molecule to break these bonds and bring about specific rearrangement of the molecules. From this point of view the action of UV radiation is simpler to understand than ionizing radiation.

The regions of the electromagnetic spectrum in which we are interested are summarized in Fig. 1. The first column gives the energy of the photon, the second column the wavelengths of the light in nm (10^{-9} meters). We are primarily concerned with UV photons in the range 200-300 nm which correspond to photons with energies between 6 and 4 ev. When it is realized that the average chemical bond is about 4 ev it can be readily understood that UV photons in this energy range can disrupt chemical bonds. In contrast, the lower part of the table shows the types of energy which are involved in x-rays. The energies of individual pho-

ENERGY (ev)	WAVE LENGTH (nm)		
1.76	700	VISIBLE	Chemical Bond ~ 4 ev
2.06	600		
2.47	500		
3.05	400		
4.06	300	U.V.	U.V. of interest in photochemistry.
6.10	200		
10,000		Low Energy X-rays	33ev/ion pair
100,000			
1,000,000		High Energy X-ray	
20,000,000			

Fig. 1. The energies of photons in the UV and X-ray regions of the electromagnetic spectrum.

tons are now very large and many different ionized species are produced when these photons interact with the biological system.

PHOTOPRODUCTS PRODUCED IN NUCLEIC ACID COMPONENTS BY ULTRAVIOLET RADIATION

The DNA and RNA of a cell contain the bases adenine, guanine, cytosine, thymine and uracil which are arranged in a specific sequence which determines the genetic code of the cell. These bases are joined to sugars which are linked together by phosphorus atoms to form the now familiar Watson-Crick model of DNA. Investigations show that the components of the DNA and RNA which absorb UV light, and are damaged by it are the bases: thymine, uracil and cytosine. The structure of uracil is shown in Fig. 2. The other compounds which we will be dealing with in this paper are thymine and orotic acid, whose structures are shown in Fig. 3. All of these materials are single ring compounds with nitrogens at the 1 and 3 positions, and a double bond between the 5 and 6 positions. The numbering is shown in Fig. 3. All of these compounds show very strong absorption in the region 260-270 nm. A typical absorption spectrum is shown in the lower half of Fig. 3. Ultraviolet irradiation in this wavelength range is very strongly absorbed by the 5-6 double bond. Irradiation usually leads to a photoproduct which lacks the 5-6 double bond producing a material which lacks absorption in the 260 to 270 nm range.

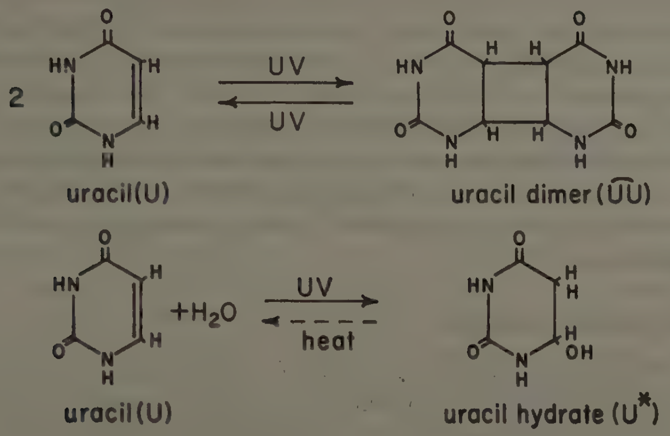


Fig. 2. Structures of the dimer and hydrate photoproducts of uracil formed by UV irradiation. The dimer is reversible to the parent uracil by re-irradiation at short wavelengths while the hydrate is reversed by heat.

Fig. 2 shows the two types of photoproducts which can be produced in these bases. When uracil is irradiated by ultraviolet light it can form a dimer shown on the right hand side of the diagram in which two of the rings are joined together by a cyclobutane ring involving bonds between the 5-5 and 6-6 positions. This photoproduct is a stable entity and has been studied extensively in the last ten years (2-7). A little thought will convince the reader that the dimer could be formed in at least four different configurations: with the rings displaced from one another as shown in the diagram or with the rings stacked on top of one another. Two more structures are possible with configurations similar to these two but with links between 5-6 and the 6-5 positions. The structures have been proved by x-ray diffraction (8, 9).

A second photoproduct which may be produced in uracil and cytosine is the hydrate. Radiation in the presence of water brings about the addition of a water molecule across the 5-6 position, leading to the hydrate which is shown on the right side of Fig. 2. This hydrate can be reversed to the parent through heat. The uracil has a strong absorption at 260 nm while the hydrate and dimer which lack the 5-6 double bond have practically no absorption. Hence, irradiation of uracil in solution will lead to a loss of absorption and this will be returned on heating of the material.

Before leaving this topic it might be worthwhile to indicate how dimers might interfere with the operation of the cell. According to the Watson-Crick model, when DNA replicates each individual strand unwinds from the double helix and a new strand is formed on each of the original strands. It can well be imagined that if, for example, two adjacent thymines in the DNA chain are linked together to form a dimer, this would prevent replication of the DNA. Some types of cells have developed methods for dealing with this type of radiation damage, and have the ability to excise these dimers before DNA replication is attempted (10, 11).

In this paper we are primarily concerned with the excited states of the bases which are the precursor to the final photoproduce—the hydrate and the dimer. The precursor state which gives rise to the final photoproducts will depend on the physical configuration of these molecules. In DNA the bases are very close together and one type of mechanism might be involved in photoproduct production in DNA, while quite a different one might be involved in the production of photoproducts of the bases in solution (12). In this paper we will be concerned with the precursor states which are responsible for photoproducts produced in the bases irradiated in solution.

Energy Level Diagram for Bases

Fig. 3 shows the absorption spectrum of uracil as a function of wavelength. The absorption band centred at 260 nm is very broad, extending from about 210 nm to about 300 nm. The bases have a large number of electrons and the molecular orbitals are very complex. Electrons are placed in these molecular orbitals in pairs with opposite spin and since the molecule has an even number of electrons the net spin $S = 0$ and the multiplicity $2S + 1 = 1$. The ground state is thus a singlet S_0. The excited state S_1 is also singlet since it is produced by the excitation of one electron without any spin change. Possible lines of excitation are shown on the top part of Fig. 3. If a photon of energy 240 nm is used to excite the molecule, the molecule is raised to a high vibrational level in the S_1 manifold. If on the other hand, a low energy photon at wavelength 280 nm is used to excite the compound it will be excited to a lower vibrational level. These vibrational levels are very close together and completely overlap so that the absorption spectrum is a continuum. If the molecule is excited to a high vibrational level it is generally

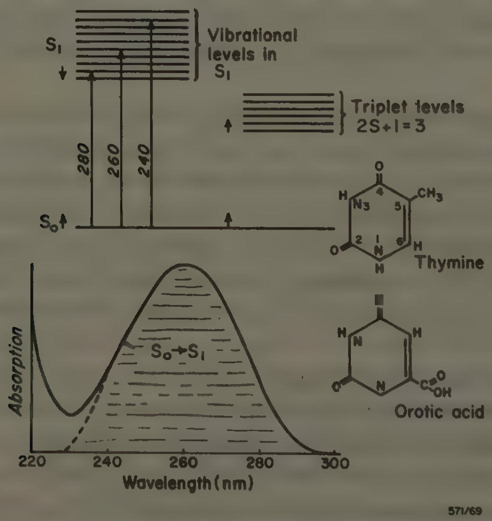


Fig. 3. Absorption spectrum and energy level diagram for uracil. Only the first excited singlet state S_1 and the corresponding triplet state T_1 are shown. The structures of thymine and orotic acid are also shown.

believed that this molecule will lose its vibrational energy before it is deexcited, so that in emission all transitions start from the lowest level in S_1. The lifetime of a molecule in the excited singlet state is very short, being of the order of 10^{-11} seconds.

We now enquire into the mechanism for dimerization. Since dimerization must involve a collision between one excited molecule and a partner in the right configuration to form a cyclobutane ring it is hard to see how this could happen in the short lifetime of the singlet state. It seems reasonable then to assume that a dimer must result from interaction between a longer lived precursor and the ground state molecule. Such a longer lived precursor is the triplet state. If we imagine that the electron which is in the excited state undergoes a spin inversion, the total molecule will have a spin of 1.0 and the multiplicity $2S + 1 = 3$ and we refer to the state as a triplet state. Presumably the triplet level has a number of higher vibrational levels as indicated in Fig. 3. The triplet level will have a long lifetime since the transition to the ground state is spin forbidden. If then, a molecule could get into a triplet state, it might live long enough to find a molecule in the ground state, with which it could combine to give the dimer. In this discussion we will show that these triplet states do exist, are responsible for dimers, and will indicate how they may be measured. On the other hand, if a hydrate is formed it requires only the collision between an excited molecule and a water molecule which are very prevalent in a solution in which the base is dissolved. It has been shown that the hydrate arises from interactions in the singlet state (13, 14). It will not be discussed further here.

Experimental Method

Two rather different experimental approaches may be used to study this problem. The first involves steady state irradiation of the material, and the detection of the photoproduct by some chemical method. This technique allows observation of photoproducts which are stable over several hours.

The second approach is to produce an intense enough flash of ultraviolet light to produce a sufficient concentration of triplet states that these can be detected by electronic methods. We will now describe these two methods in detail.

Steady State Irradiation: A schematic arrangement for steady state irradiation is shown in Fig. 4. Light from a large UV monochromator (15, 16, 17) which is not shown emerges from the exit slit, passes through a slit lens and is focussed by a large quartz aplanatic lens into a quartz cuvette (dimensions 1 cm × 1 cm × 3 cm). The beam of UV light is monitored by a photocell which is arranged so that it can be moved laterally to intercept the beam. At the same time, the quartz cuvette is moved on to the axis of a spectrophotometer which may be used to measure the absorbance of the solution. With such an arrangement it is possible to carry out irradiations with accurate dosimetry so that the amount of irradiation which each molecule in the vessel sees can be calculated. The unit of exposure in this type of work (16) is usually given as photons per cm^2 or micro-Einstein per cm^2. A micro-Einstein (μE) is a micro mole of photons, that is 6.02×10^{17} photons. So, an exposure of 1 $\mu E/cm^2$ is 6.02×10^{17} photons per cm^2.

Samples of the solution containing radioactive thymine are removed from the quartz cuvette during the irradiation and are spotted on chromatographic paper. When suitable solvents flow over this paper, the various photoproducts move at different speeds down the paper giving the separation shown in Fig. 5 (17, 18).

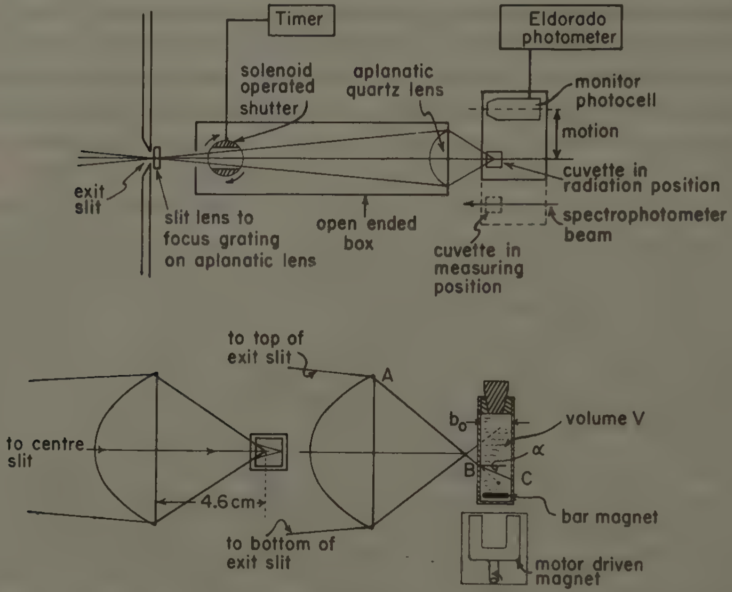

Timer
Eldorado photometer
solenoid operated shutter
aplanatic quartz lens
monitor photocell
motion
cuvette in radiation position
exit slit
slit lens to focus grating on aplanatic lens
open ended box
spectrophotometer beam
cuvette in measuring position
to top of exit slit
A
b_0
volume V
α
to centre slit
B
C
4.6 cm
bar magnet
to bottom of exit slit
motor driven magnet

Fig. 4. Details of components at the exit slit of monochromator.

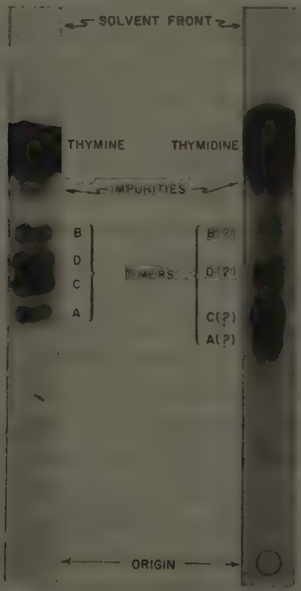

SOLVENT FRONT
THYMINE
THYMIDINE
IMPURITIES
B
D
C
A
DIMERS
B(?)
D(?)
C(?)
A(?)
ORIGIN

Fig. 5. Typical separation of the four dimers of thymine from each other and from the parent by thin layer chromatography (18).

The four dimers of thymine are well separated from the parent molecule. The position of the spots is determined by exposing the radioactive chromatograph to x-ray film. The amount of photoproduct in each of these spots can then be determined by cutting out the spot and counting it in a liquid scintillation counter. The yield of each of the photoproducts as a function of exposure can be determined. A typical set of kinetic data for orotic acid (19) is shown in Fig. 6. The insert shows the percentage activity in the dimer spot (in this case there is only dimer) as a function of exposure in $\mu E/cm^2$. Over the small dose range shown in the insert, this gives a straight line whose slope is a measure of the probability or cross section for the production of this dimer. These cross sections can be expressed in $cm^2/\mu E$. The cross section used in this way is exactly analogous to the cross section used by the nuclear physicist in talking about the activation of material in a pile.

In the case of orotic acid, only one dimer is formed and with the formation of each dimer, the solution loses some absorbance, so that if the percentage loss of absorbance is plotted against exposure, the curve in the main part of Fig. 6 is

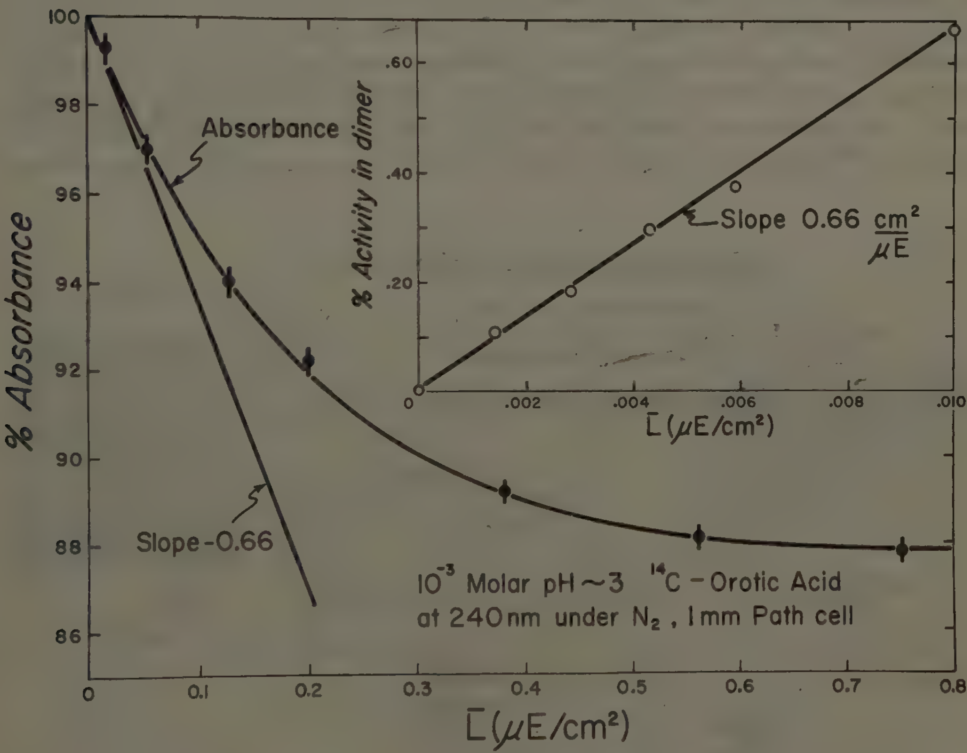


Fig. 6. Measurement of initial rate of dimer formulation by two techniques. Exposure scale of the main figure is 50 × that of the insert (19).

obtained. This curve quickly reaches a saturation value when about 12% of the absorbance has been lost. This saturation is due to two competing reactions, a forward reaction which leads to the dimer, and a reverse reaction which leads to the monomer. The dimers can be converted back to the monomer by the same radiation which produces them. It is essential then to determine the initial slope of such a graph. Methods for doing this are described in detail elsewhere (14, 17) but it is obvious from Fig. 6 that one good way is to measure the activity in the dimer spot and to study the yield of this dimer at very small exposures.

Flash Photolysis: The technique of flash photolysis (20, 21, 22) is illustrated schematically in Fig. 7. The solution which is to be examined is held in a quartz cylindrical tube and on either side of it are placed two flash lamps. These consist of quartz tubes with solid quartz rods down their centre leaving an annular space in which the plasma is formed. When the large condenser is discharged through the switch *S*, the flash tubes discharge, producing the transient species in the central cylindrical tube. The transient species is observed by passing light from the analyzing lamp *L* through this solution to a monochromator *M*, to a photomultiplier *P*, the output of which is fed to a cathode ray oscilloscope. Rather complex electronics is required to control such a system, and this is indicated schematically at the bottom of Fig. 7. When a picture is required the lamp *L* is brightened by a DC pulse which lasts for about 20 msec. After the lamp reaches a steady brightened state, the scope is triggered and immediately afterwards the switch *S* is closed to discharge the condensers through the flash lamps, producing the transient species in the observation vessel. Species can be observed over a period from 1 to 200 μsec. The transient presented on the scope is photographed by polaroid film.

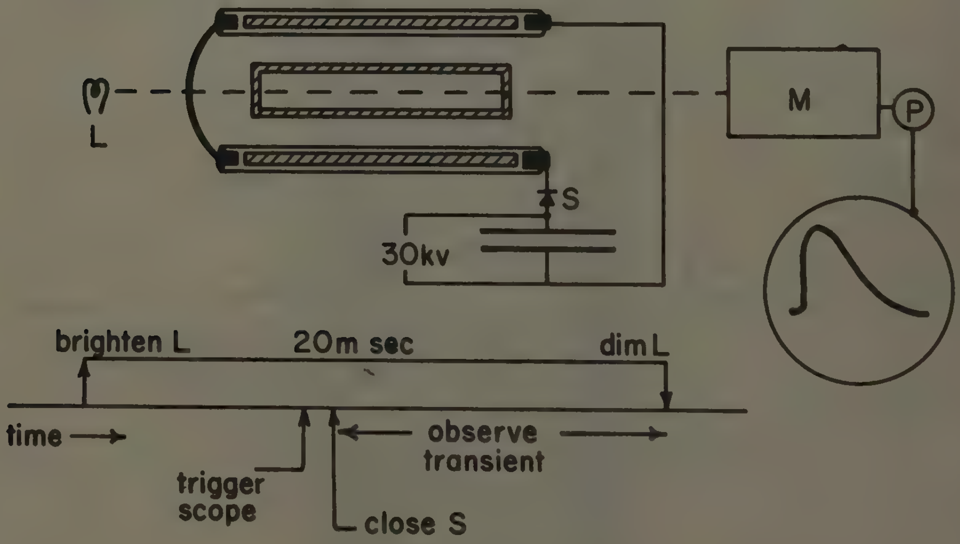
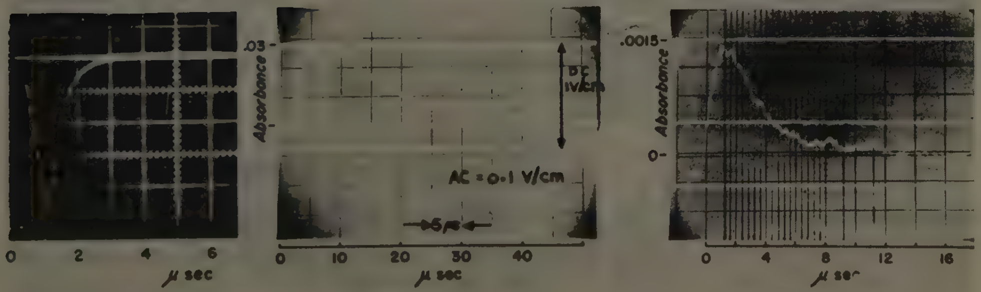


Fig. 7. Schematic diagram for the flash photolysis apparatus

A typical trace with water in the reaction vessel and *L* turned off is shown in Fig. 8a. This trace shows the intensity of light from the flash tubes as a function of time. It is seen that the flash is essentially over in 3 μsec. This very rapid termination of the flash was achieved by placing a quartz rod down the axis of the flash lamp so that all molecules of plasma have a very short path to the nearest quartz thus quenching the flash rapidly. A short flash is essential if short lived transient species are to be observed.

If a reactive species such as a triplet is produced in the solution by the flash it can be detected from the change in the light level reaching the photomultiplier from *L*. Such a transient is shown in Fig. 8b. This transient shows the absorption of the solution as a function of time. (Positive values now correspond to less light reaching the photomultiplier while in Fig. 8a positive values mean more light.) Initially this absorbance increases as the species is formed by the flash—then the species decays with time. If one plots the absorbance as a function of time on semilog paper a straight line is obtained from which the mean lifetime may be calculated. It should be emphasized that the transient is observed by removing photons from the beam. If the absorbing species is a triplet then this absorption is due to absorption of photons by the triplet state which will raise some molecules from T_1 (see Fig. 3) to higher levels such as T_2 or T_3 in the triplet manifold. The levels T_2 and T_3 are not shown in Fig. 3. The transient absorbing species shown in Fig. 8a has its peak absorption at about 320 nm and probably corresponds to a T_1 to T_3 absorption although this has not been rigorously proven.



100 joules
~1μE
Herbert
and
Whillans
april 69

8.2×10^{-5} M
Orotic acid
pH-5
330 nm
Herbert
april 23/69

10^{-4} M uracil
N_2 –
τ = 2.5μ sec
350 nm
Whillans
april 21/69

Fig. 8a. Typical flash from quartz lamps; 8b. Transient species observed in orotic acid; 8c. Species in uracil.

REACTION MECHANISM

The reaction mechanism which accounts for our observed data is given in Fig. 9. This scheme was proposed initially by Sztumpf-Kulikowska, *et al.* (23). The absorption of a photon leads to excitation into the singlet manifold. The absorption cross section is σ_a, expressed in cm$^2/\mu$ mole. The number of singlet states then formed by an exposure dL expressed in μE/cm^2 is given by

$$dS = \sigma_a[P] \cdot dL \tag{1}$$

where $[P]$ is the concentration of the parent material.
The singlet state can lose its energy by a number of competing pathways. It may lose its energy through radiationless transitions to the ground state. These radiationless transitions are not well understood but are presumably due to interactions between the excited state and the solvent. It could lose its energy by emission of radiation (fluorescence) or by reacting with the solvent to form some type of photoproduct such as the hydrate. It could also suffer a spin inversion leading to a triplet state. The quantum yield for this will be represented by ϕ_{isc}. The quantum yield ϕ_{isc} is the number of triplet states produced per absorbed photon.

The number of triplet states dT arising from the excitation of the singlet manifold is given by

$$dT = \phi_{isc} \cdot \sigma_a \cdot [P] \cdot dL \tag{2}$$

We now enquire into methods by which these triplet states can lose their energy. A number of possible mechanisms immediately come to mind. The triplet state could lose its energy by phosphorescence, by a radiationless transition leading to the ground state, it could be quenched by oxygen, which can supply the unpaired spin leading to the ground state, or it could combine with a molecule in the ground state leading to a dimer. Since phosphorescence at room temperature is not observed, we only need consider the last three pathways.

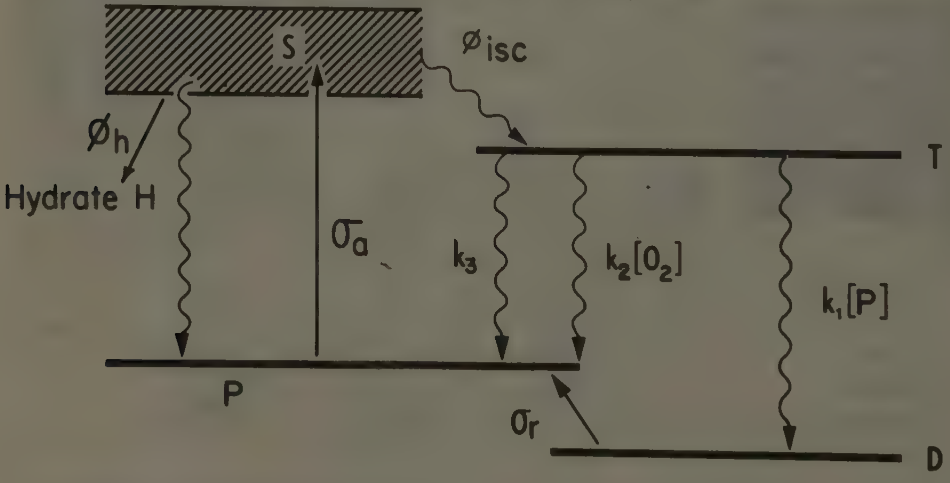


Fig. 9. Reaction scheme for the dimerization and hydration of uracil in acqueous solution.

Let T represent the concentration of triplet states (moles/liter) at any given time. Then the number of triplet states which go to the ground state per sec by radiationless transitions is $k_3[T]$, where k_3 is the rate constant with dimensions sec^{-1} and $k_3[T]$ has dimensions moles liter^{-1} sec^{-1}.

Similarly, let $k_2[O_2]\,[T]$ be the number of triplet states which are quenched by oxygen per sec. k_2 is a bimolecular rate constant with dimensions sec^{-1} moles^{-1} liter.

Finally, let $k_1[P]\cdot[T]$ be the number of triplet states per sec which are quenched by collisions with the parent molecule at concentration $[P]$, and which lead to dimers. k_1 is also a bimolecular rate constant with dimensions sec^{-1} moles^{-1} liter.

Combining these we have

$$\frac{dT}{dt} = T\{k_3 + k_2[O_2] + k_1[P]\} \tag{3}$$

which may be integrated to give

$$T = T_o\, e^{-t/\tau} \tag{4}$$

where T_o is the initial concentration of triplet states, τ is the mean lifetime and $1/\tau$ is given by

$$\frac{1}{\tau} = (k_3 + k_2[O_2] + k_1[P]) \tag{5}$$

This equation shows that the lifetime of the triplet state depends on oxygen concentration and on the concentration of the parent. The higher the oxygen concentration and the higher the parent concentration, the shorter the lifetime.

These equations show that the population of the absorbing species (the triplet state) should decrease exponentially with time. Further equation (5) shows that a plot of $1/\tau$ versus either $[O_2]$ or $[P]$ should be linear and enable us to determine the three rate constants k_1, k_2 and k_3.

If we are interested in the results of steady state irradiation we wish to determine the fraction of the triplet states which lead to dimers. Clearly, the fraction of the states which go to dimers is given by

$$f = \frac{k_1[P]}{k_1[P] + k_2[O_2] + k_3} \tag{6}$$

since the numerator is the pathway leading to the dimers, and the denominator is the sum of all the pathways leading to the quenching of the triplet state. If we now combine equation 2 and 6, we determine the number of dimers which will be produced by an increment of exposure dL.

$$dD = \sigma_a[P]\cdot\phi_{isc}\cdot\frac{k_1[P]}{k_1[P] + k_2[O_2] + k_3}\cdot dL \tag{7}$$

It is convenient to rearrange this equation in the form of 8*a* and 8*b*.

$$\frac{1}{R} = \left[\frac{[P]}{dD/dL}\right]_o = \frac{1}{\sigma_a\phi_{isc}}\left\{1 + \frac{k_2[O_2]}{k_1[P]} + \frac{k_3}{k_1[P]}\right\} \tag{8}$$

In this equation the subscript on the square bracket merely indicates that the ratio is to be evaluated at zero exposure. This avoids the problem resulting from the reversal of the dimers discussed earlier in connection with Fig. 6. This equation shows that if $\frac{1}{R}$ is plotted against P^{-1} under conditions of zero oxygen pressure, a straight line should be obtained which will enable us to determine σ_a. ϕ_{isc} and the ratio k_3/k_1 On the other hand, if at a fixed concentration of the parent material, $1/R$ is plotted against the oxygen concentration, then it should be possible to determine k_2/k_1.

Results from Steady State Irradiation: Typical results showing a plot of $1/R$ versus $[P]^{-1}$ and $[O_2]$ are given in Fig. 10. This shows that straight lines are obtained from which the ratio of rate constants may be calculated. However, no manipulation of the steady state data can yield absolute rate constants but only their ratios.

Results from Flash Photolysis

Typical results from the flash photolysis of uracil (22) are given in Fig. 11. Here $1/\tau$ is plotted against uracil concentration and $[O_2]$ concentration yielding good straight lines. The intercept from Fig. 11a gives (see equation 5) k_3. The

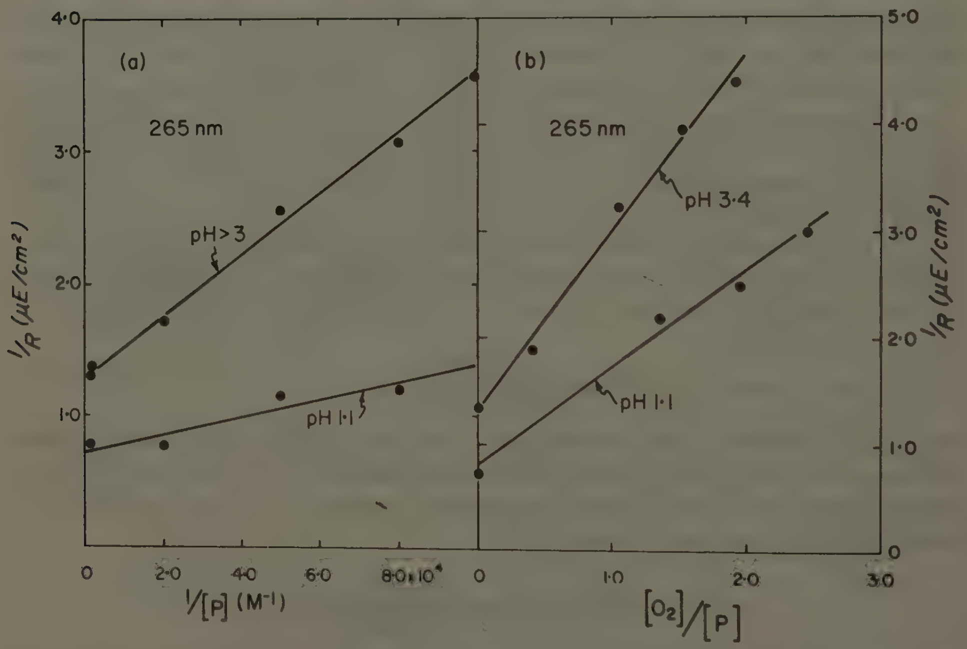


Fig. 10. Plot of $1/R$ at 265 nm vs.: (a) reciprocal of orotic acid concentration, $[P]^{-1}$, for neutral (pH 1 • 1) and ionized (pH > 3) forms, $[O_2] = 0$; (b) ratio of oxygen concentration to orotic acid concentration, $[O_2]/[P]$ for the same two forms. (19).

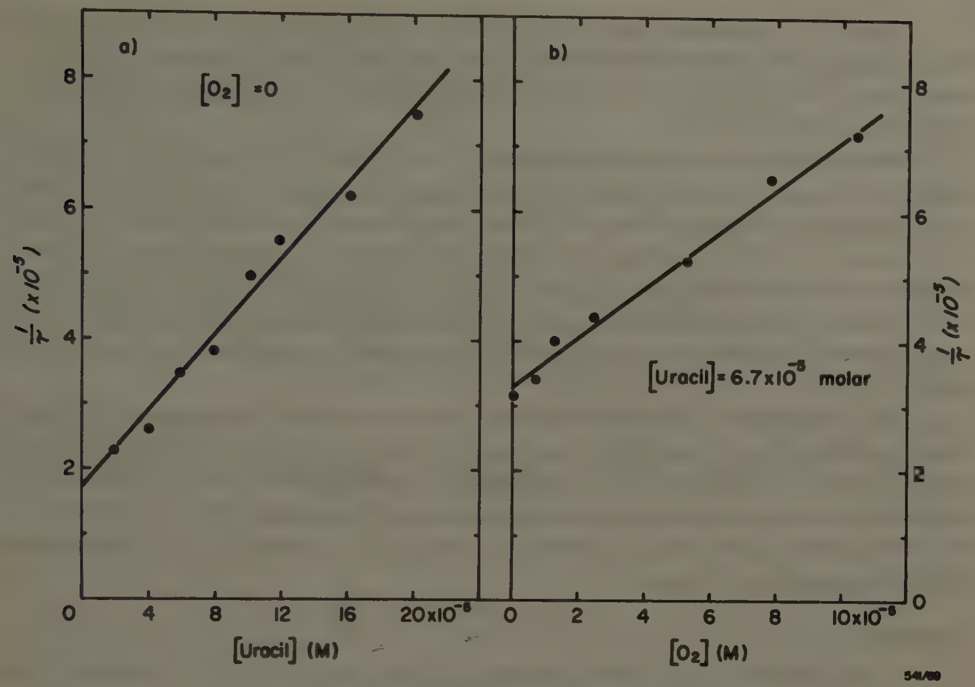


Fig. 11. Studies in quenching of the uracil triplet by
(a) other molecules of uracil
(b) molecular O_2 dissolved in solution.

quantity $1/k_3$ is the lifetime τ_0 of the triplet state when it is unquenched by either oxygen or parent, that is in very dilute solutions. For uracil this lifetime is 6.1 μsec. Under any actual conditions the lifetime will be shorter than this.

The ratio of the rate constants can also be obtained from the flash experiments. These agree within experimental error with our steady state irradiations. This proves that the transient species observed in flash is the one responsible for dimer production and that dimers do come from the triplet state. The rate constants for uracil and orotic acid are given in Table 2 (14, 19, 21, 22).

Table 2. Rate Constants for the Quenching of the Triplet State in Uracil and Orotic Acid.

	$\tau_0 = k_3^{-1}$ (μsec)	k_3 (sec^{-1})	k_2 (M^{-1} sec^{-1})	k_1 (M^{-1} sec^{-1})
Uracil	6.1	16×10^4	4.1×10^9	2.9×10^9
Orotic acid				
unionized (pH 1.1)	110	0.91×10^4	2.0×10^9	1.8×10^9
ionized (pH > 3)	80	1.25×10^4	0.81×10^9	0.68×10^9

Tentative results have also been obtained for thymine (18). The intersystem crossing is smaller than in either uracil or orotic acid so the observed transient is very weak. The lifetime of the thymine triplet under unquenched conditions is about the same as in uracil.

Many questions are still unanswered about these rate constants. Why for example, is τ_0 smaller by a factor of 10 in uracil as compared to orotic acid? Simlarly, why are the rate constants k_2 for oxygen quenching so different for the species given in Table 2? Also, why does k_1 the rate constant for a collision between an excited state and a ground state molecule depend so much on the species? Intensive investigations are now under way to see how these rate constants depend upon the groups which may be placed on the pyrimidine rings (see Fig. 2 and 3).

The properties of the observed triplet state need more investigations. We still need to determine its absorption spectrum precisely and the extinction coefficient of this triplet. It is hoped that investigations of this kind will yield a good deal of information concerning the spectroscopic nature of the energy levels in the bases of DNA. Such information is essential if we are to understand the mechanisms underlying radiation damage.

Summary

1. Experiments have been described which enable one to determine the amount of dimer produced in DNA bases in solution by UV as a function of concentration of the parent and of oxygen.

2. Experiments have also been described to indicate how triplet states of the bases may be detected in flash photolysis.

3. The lifetime of the triplet state has been measured under different conditions. It is quenched with the same concentration dependence as is dimer production proving that dimers in solution come from the triplet state.

Acknowledgements

The author takes pleasure in acknowledging the continued financial assistance of the Medical Research Council of Canada and the National Cancer Institute of Canada.

REFERENCES

1. E. J. Hart, Radiation Chemistry, Vol. 1, Advances in Chemistry Series 81, Washington 1968.
2. K. C. Smith in Photophysiology (A. C. Giese ed.) Vol. 2, 329, Academic Press, New York, 1964.
3. K. C. Smith, *Rad. Res. Supplement* **6,** 54, 1966.
4. J. K. Setlow in Current Topics in Radiation Research (M. Ebert and A. Howard, eds.) Vol. 2, 195, North Holland Publ. Amsterdam 1966.
5. R. B. Setlow, *J. Cell. Comp. Physiol.* **64,** (Suppl 1) 51, 1964.
6. R. Beukers and W. Berends, *Biochim. Biophysica Acta* **49,** 181, 1961.
7. H. E. Johns, Third Int. Congress Rad. Res. 1966, North Holland Publishing Co., Amsterdam, 1967.

8. D. Weinblum and H. E. Johns, *Biochim. Biophys. Acta* **114,** 450, 1966.
9. N. Camerman, S. C. Nyburg and D. Weinblum, *Tetrahedron Letters* **42**, 4127, 1967.
10. R. P. Boyce and P. Howard-Flanders, *Proc. Natl. Acad. Sci. U.S.* **51,** 293, 1964.
11. R. B. Setlow and W. L. Carrier, *Proc. Natl. Acad. Sci. U.S.* **51,** 226, 1964.
12. J. Eisinger and R. G. Shulman, *Science* **161,** 1311, 1968.
13. C. L. Greenstock, I. H. Brown, J. W. Hunt and H. E. Johns, *Biochem. Biophys. Res. Comm.* **27,** 431, 1967.
14. I. H. Brown and H. E. Johns, *Photochem. Photobiol.* **8,** 273, 1968.
15. H. E. Johns and A. M. Rauth, *Photochem. Photobiol.* **4,** 673, 1965, also **4,** 693, 1965.
16. H. E. Johns, *Photochem. Photobiol.* **8,** 547, 1968.
17. H. E. Johns, Methods in Enzymology, 1969.
18. G. J. Fisher and H. E. Johns, *Photochem. and Photobiol.* 1970.
19. D. Whillans and H. E. Johns, *Photochem. Photobiol.* **9,** 323, 1969.
20. J. W. Boag, *Photochem. and Photobiol.* **8,** 565, 1968.
21. M. A. Herbert, J. W. Hunt and H. E. Johns, *Biochem. Biophys. Res. Comm.* **33,** 643, 1968.
22. D. W. Whillans, M. A. Herbert, J. W. Hunt and H. E. Johns, *Biochem. Biophys. Res. Comm.* **36,** 912, 1969.
23. E. Sztumpf-Kulikowska, D. Shugar and J. W. Boag, *Photochem. Photobiol.* **6,** 41, 1967.

PHYSICAL LIMITS OF MANNED UNDERWATER ACTIVITY

H. R. Schreiner

I wish to thank you for inviting me to come here today to talk about one of the most exciting new frontiers of medical physics. The State of Massachusetts and its historic capital of Boston, whose development has been so uniquely shaped by the sea, seem a most appropriate setting for this presentation. It was a native son of this state, John F. Kennedy, who focussed the attention of our nation on the importance of the sea when he said, eight years ago,

> "We are just at the threshold of our knowledge of the oceans. Already their military importance, their potential use for weather prediction, for food, and for minerals are evident. Knowledge of the oceans is more than a matter of curiosity. Our very survival may hinge upon it."

Never before in the history of man has the sea held greater promise for the fulfillment of his needs than today. Since the beginning of this decade there has been a remarkable and rapidly accelerating acceptance of the commercial and military importance of the sea in general and of the continental shelves in particular (75). This in turn has brought with it a revitalized and greatly intensified interest in problems associated with man's entry into the sea.

Most of these problems are created by the impact of the physical facts of submergence on the biological peculiarities of man.

The physical facts of submergence are determined by the physical properties of seawater (Fig. 1), especially its temperature and thermal conductivity, its refraction and absorption of light, its salt content and wetting ability, its density which affects the buoyant weight of objects submerged in it, and, last but not least, its mass which is responsible for the pressure which a column of seawater exerts.*

---

* In this discussion the pressure exerted by 1 foot of seawater (fsw) is held to equal 1/33 of one atmosphere or 23.033 mm Hg.

While heat loss (6), subsea vision (43), weakening of the skin structure, and the reduced relative weight of the submerged body represent important diving problems, it is the inability of the human lung to function properly in the face of an uncompensated increase in external pressure which produces the most significant limitation of manned underwater activity. This is so because the required compensation of the pressure produced by the water column—which equals 760 mm Hg for every 33 feet of vertical distance**—on the diver creates a veritable Pandora's box of complex and interrelated secondary problems (Fig. 2).

Seawater

Physical Properties	Biological Significance
Temperature and Thermal Conductivity	Heat Loss
Light Refraction and Absorption	Visual Problems
Salt Content and Surface Tension	Weakening of Skin Structure
Specific Gravity	Near Weightlessness
Mass	Pressure

Fig. 1

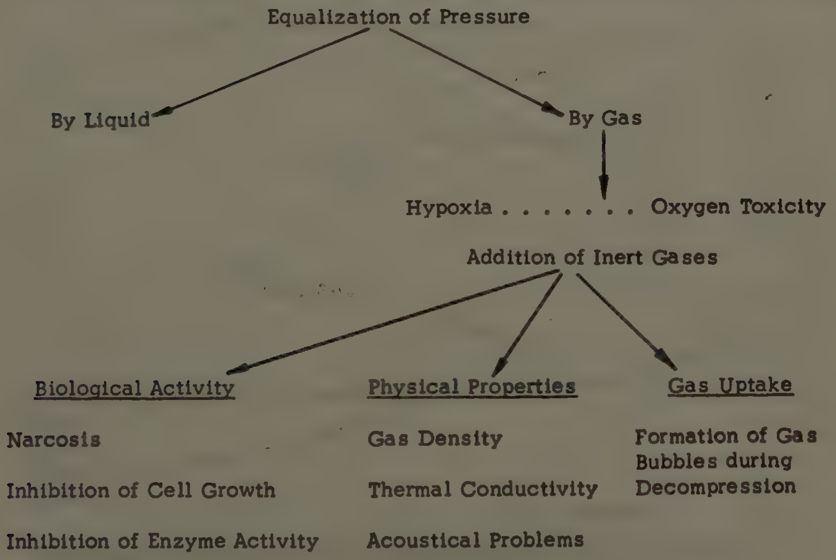

Equalization of Pressure
By Liquid
By Gas
Hypoxia . . . . Oxygen Toxicity
Addition of Inert Gases
Biological Activity
Narcosis
Inhibition of Cell Growth
Inhibition of Enzyme Activity
Physical Properties
Gas Density
Thermal Conductivity
Acoustical Problems
Gas Uptake
Formation of Gas Bubbles during Decompression

Fig. 2

** Assuming a density of 1.027.

In theory, any inert fluid, gas or liquid, may be employed to fill the lungs to assure their continued ability to function under pressure. The use of oxygenated physiological saline or of fluorocarbon fluids as respiratory media has been shown to be feasible in small and large mammals (46) and may, in years to come, find application in manned diving as well. For the time being, however, there is no substitute for the breathing of gas.

Pure oxygen, breathed under pressure, produces pulmonary damage and functional abnormalities of the central nervous system. These toxic effects of oxygen are time and pressure dependent. Since the limits (47) of these effects are reasonably well known (Fig. 3), they can be avoided by diluting with an inert gas the oxygen that must be supplied to the diver. Maintaining an adequate degree of this dilution can pose difficulties. For example, for prolonged underwater exposures, the difference between the non-hypoxic minimum and the non-toxic maximum concentration of oxygen is in the order of only 0.3 atm (Fig. 4) which, especially at great depth and for the free-swimming diver, makes extraordinary demands on the precision and reliability of oxygen-sensing and control instrumentation.

The selection of an inert diluent for oxygen in deep diving breathing gas mixtures must be regarded from three separate points of view: (A) the known biological activity of all so-called inert gases, (B) the changes in physical properties of the gaseous environment brought about by the introduction of the diluent, and (C) the uptake of the diluent by the tissues of the body.

It has been more than one hundred years since the narcotic action of compressed air was first described (7). The euphoria, retardation of higher mental processes and impairment of neuromuscular coordination observed while breathing compressed air resemble the effect of alcohol on the central nervous system. Divers

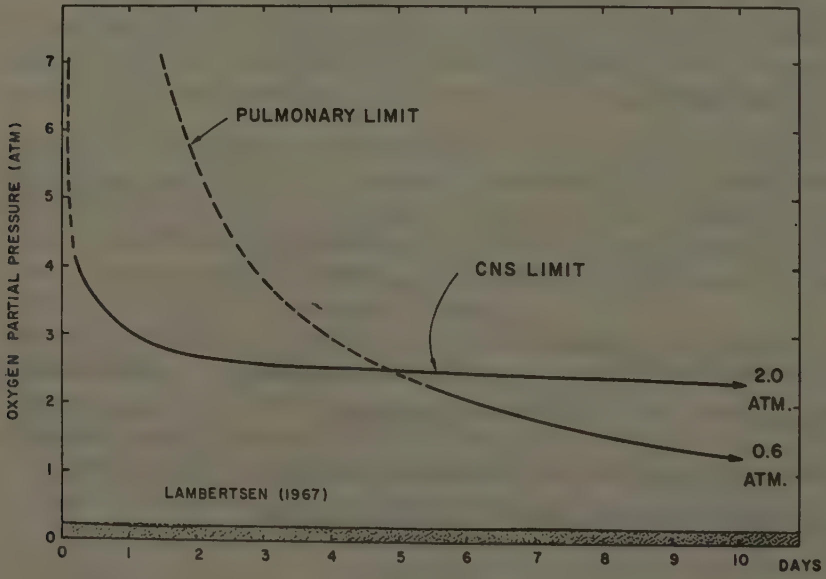


Fig. 3

who are familiar with this phenomenon speak with some justification of the existence of "Martini's Law" which holds that the effect of breathing air equals that of one dry martini for every 50 feet of descent. Objective measurements of the performance decrement seen in human subjects working in compressed air confirms this subjective evaluation. Figure 5 shows how manual dexterity is affected by compressed air (1). It also documents how exercise potentiates this effect. Intel-

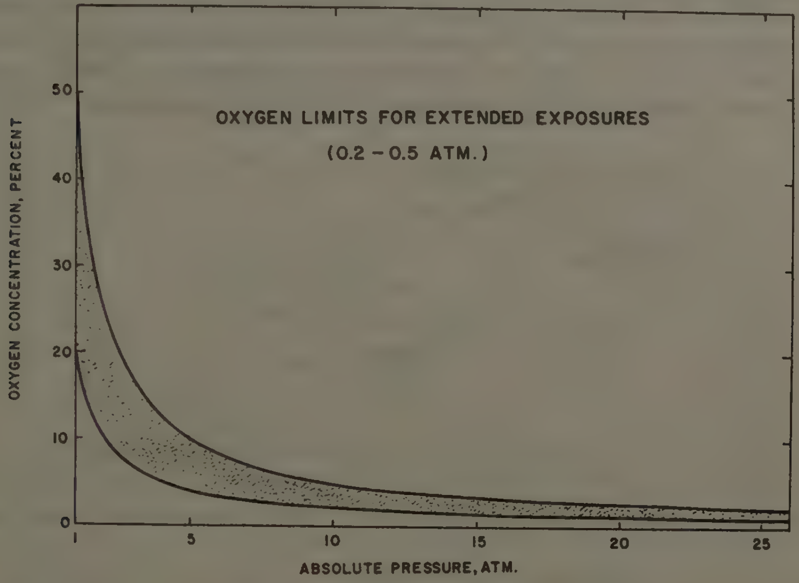


Fig. 4

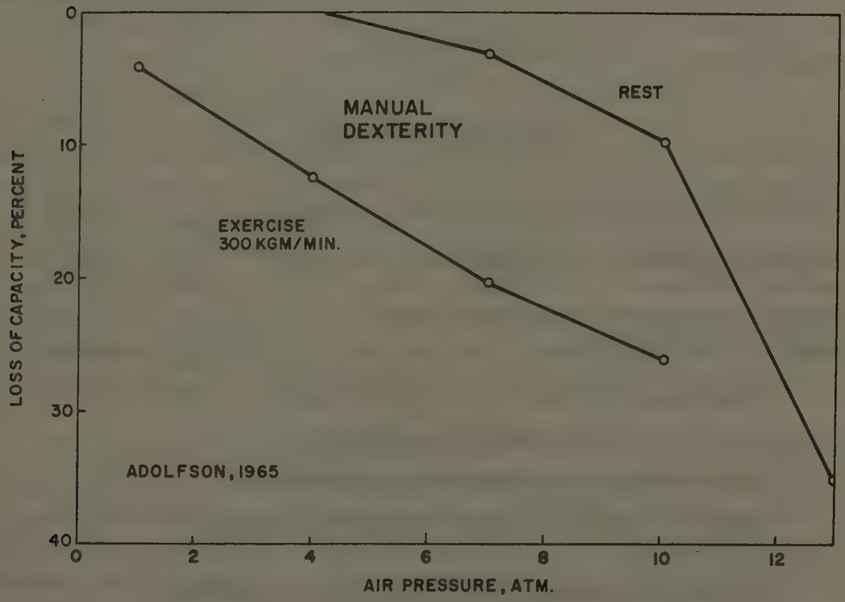


Fig. 5

lectual functions are also progressively impaired (Fig. 6) with increasing air pressure. Here, however, the potentiating effect of exercise is not nearly as important as it is on the use of simple motor skills under otherwise identical conditions. Compressed air narcosis is apparently the result of a molecular-level physical interaction of nitrogen with structures within the central nervous system. This effect of nitrogen can already be detected in man at a partial pressure of only 1.6 atm. (Fig. 7). Nitrogen narcosis (9) increases with the complexity of an attempted task, is rapid in onset and readily reversible and affects the capacity for learning to a greater extent than it inhibits the execution of a routine task. Finally, as in the case of alcohol, there are wide variations in individual susceptibility to nitrogen narcosis.

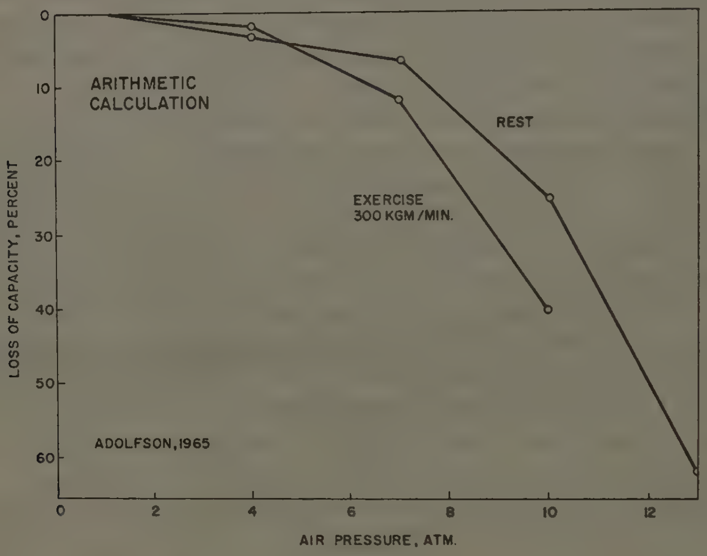


Fig. 6

NITROGEN NARCOSIS

First measureable impairment of mental performance: P_{N_2} = 1.6 atm.

Significant decrement in psychomotor performance: P_{N_2} = 3.2 atm.

Maximum decrement consistent with safe underwater performance: P_{N_2} = 5.6 atm.

Effect increases with complexity of task.

Onset is rapid and readily reversible.

Large differences in individual susceptibility.

Exercise potentiates effect especially on manual dexterity.

Capacity for learning more susceptible to inhibition than execution of routine task.

Fig. 7

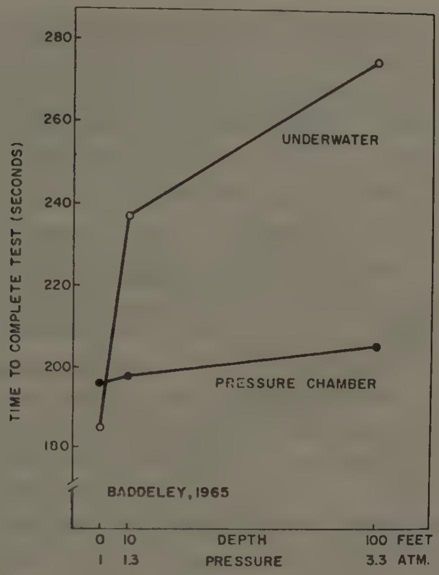


Fig. 8

Submergence represents an additional potentiating factor (5) of nitrogen narcosis (Fig. 8), a fact which must be clearly kept in mind when judging expected human performance under water by the results of measurements performed in dry chambers.

Nitrogen is by no means unique in the latent pharmacological activity which it possesses. All inert gases (69), in fact, are today known to produce measurable physiological derangements if their concentration is sufficiently high. While helium at a partial pressure of 35 atm (Bühlmann, personal communication, 1969) and neon at 16 atm (66) do not produce an unequivocal impairment of central nervous system function in man, all so-called inert gases in sufficiently high concentration will, for example, inhibit the growth of human cells in culture. Figure 9 shows data on the growth of HeLa cells in various atmospheres as a function of pressure. This data which comes from our laboratory (13, 63) is presented here in arbitrarily linearized form to accentuate the quantitative differences that are observed in the specific inhibitory activity of these gases. These differences demonstrate that the observed effects cannot be ascribed to pressure as such. With the least inhibitory gases such as helium, however, it is not possible at this time to rule out completely the possible effect of pressure per se. Qualitatively similar inhibitory effects of "inert" gases have also been observed with a number of purified enzymes (20, 21).

It is essentially impossible to introduce "inert" gases into a living system without at the same time changing the physical properties of the gaseous environment such as its density and thermal conductivity. It is therefore often quite difficult, when studying the intact mammalian organism in abnormal gaseous environments, to differentiate between pharmacological and physical effects of the inert gases. A case in point is the well-known increased heat loss (62) and metabolic stimulation which helium produces. On cursory examination, this stimulation could have a basis that is pharmacological, physical, or both. Closer review of respiratory data obtained in our laboratory (27) in various gaseous environments (Fig. 10) and at different total pressures reveals a convincing correlation between oxygen consumption and the relative convective heat loss into these gaseous environments which suggests that there is no pharmacological basis for the observed effect.

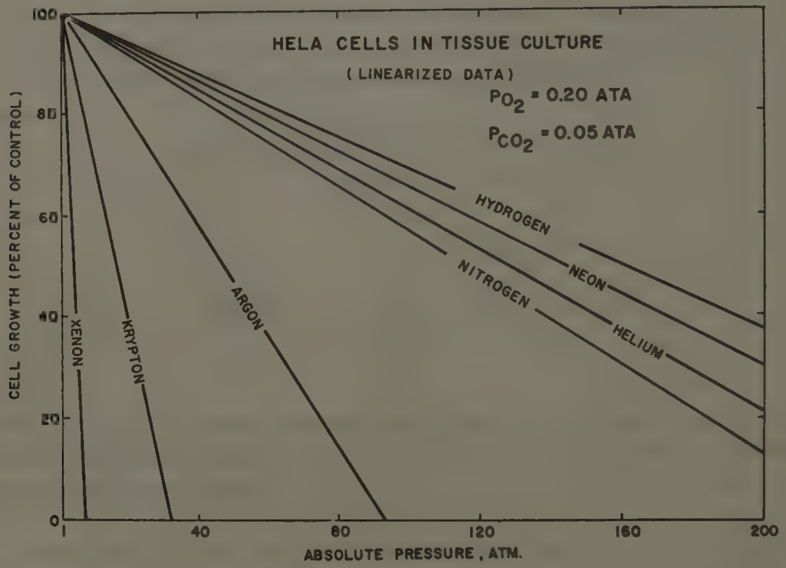


Fig. 9

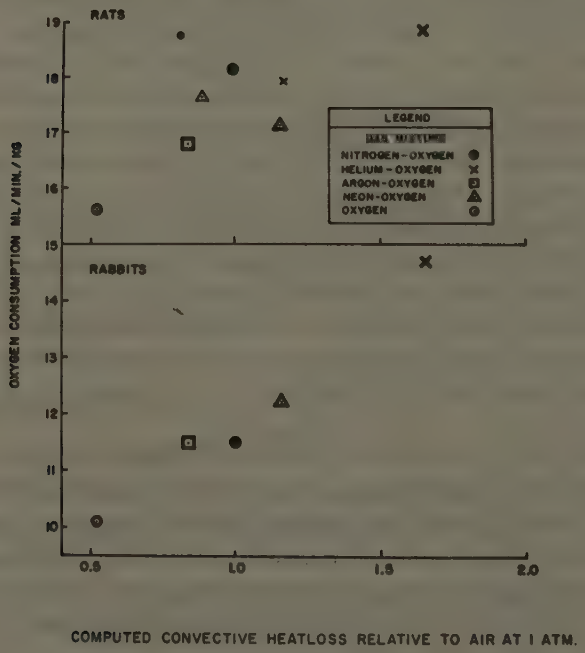


Fig. 10

The density of a breathing mixture has a significant effect on pulmonary function. Maximum voluntary ventilation, for example, drops sharply with increasing pressure (50); in a helium-oxygen mixture at 21.7 atm. abs. (with a density 3.8 times that of air at sea level) we measured (28) a 60 per cent decrease in this parameter.

A purely physical effect of certain physiologically inert gases such as helium and hydrogen is the distortion of the human voice (32). While this acoustical problem can be solved electronically (26), helium speech unscramblers are still in their experimental stage, and we must consider helium speech as a definite physical limit of manned underwater activity.

It has recently been suggested (45) that inert gases entering body tissues during compression exert osmotic pressure. Water could therefore be expected to flow along the resulting osmotic gradients which in turn may account for some hitherto unexplained disturbing and painful symptoms occasionally reported by divers undergoing compression at a rapid rate.

Another physical aspect of gases dissolving in biological tissues under pressure is the accompanying change in solvent volume (73) which in turn may produce physiological effects.

Increased gas density (53), submersion (2), respiratory incumbrances caused by mask breathing and physical exertion under diving conditions (3, 29) all tend to increase alveolar carbon dioxide levels (49) (Fig. 11). This increase in turn leads to peripheral vasodilation. Vasodilation, irrespective of the manner in which it is produced, tends to increase the tissue perfusion with blood and hence the rate of inert gas transport to and from the tissues. As we will see later on, accelerated removal of inert gases from the tissues is indeed desirable during decompression but during exposure of the diver to maximum pressure an enhanced uptake of

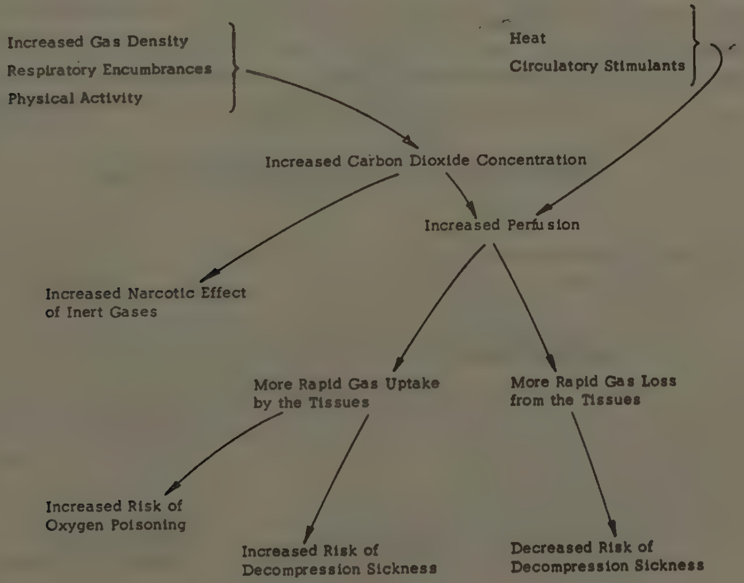

Increased Gas Density
Respiratory Encumbrances
Physical Activity
Heat
Circulatory Stimulants
Increased Carbon Dioxide Concentration
Increased Perfusion
Increased Narcotic Effect of Inert Gases
More Rapid Gas Uptake by the Tissues
More Rapid Gas Loss from the Tissues
Increased Risk of Oxygen Poisoning
Increased Risk of Decompression Sickness
Decreased Risk of Decompression Sickness

Fig. 11

inert gas increases the time required to return him safely to surface pressure. Another aspect of this interaction of pharmacological, physiological, and physical parameters is the enhanced susceptibility to oxygen poisoning which is brought on by an increase in the rate of perfusion.

The return of the diver from great depths must be accomplished safely and with reasonable speed. Inert gases that have been transported to the peripheral tissues by the circulatory system must be eliminated via the circulatory system without coming out of solution before reaching the alveolar-capillary membrane. As we will shortly see, the rate of inert gas transport is in the main dependent upon the magnitude of the inert gas partial pressure gradient between the alveolar spaces and the peripheral tissues. It is therefore tempting to employ a high level of oxygen during the bottom phase of a dive to minimize the uptake of inert gas, and during decompression to hasten inert gas elimination. This course of action, however, exacts its physiological price (Fig. 12) because oxygen, owing to its vasoconstrictive effect, reduces the rate of perfusion (33). While this may be desirable under maximum pressure, during decompression it partially nullifies the desired enhancement of gas elimination. Other means of producing vasoconstriction yield qualitatively similar results. The exposure of the diver to high tensions of oxygen moreover enhances the risk of oxygen poisoning. While attitudes on this point vary widely, it is unlikely that central nervous system effects of oxygen would occur at an oxygen partial pressure of 1.2 atm. or less. Finally, there are some grounds (23) to believe that increased levels of oxygen enhance either directly or indirectly the narcotic effect of inert gases.

Traditional concepts of decompression theory which have served well to return divers from depths to about 300 feet have failed, sometimes spectacularly, to provide guidance to adequate decompression from considerably greater depths.

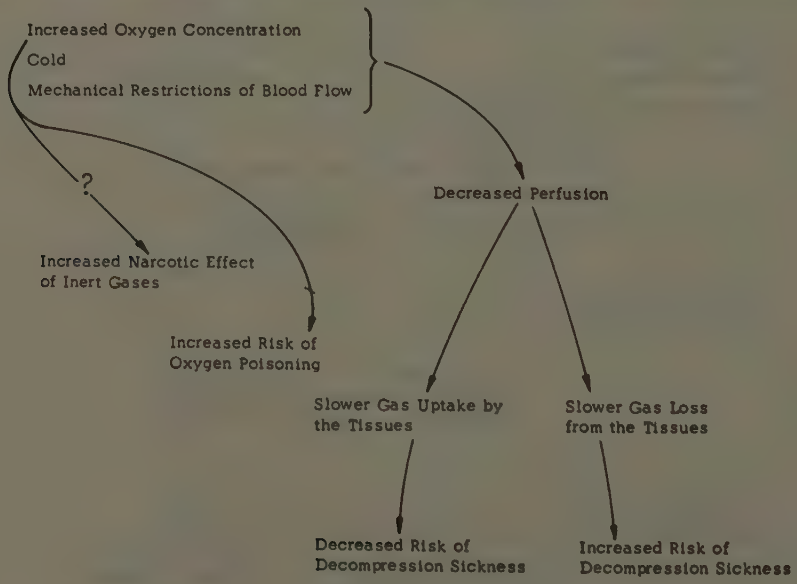

Increased Oxygen Concentration
Cold
Mechanical Restrictions of Blood Flow
?
Decreased Perfusion
Increased Narcotic Effect of Inert Gases
Increased Risk of Oxygen Poisoning
Slower Gas Uptake by the Tissues
Slower Gas Loss from the Tissues
Decreased Risk of Decompression Sickness
Increased Risk of Decompression Sickness

Fig. 12

This has necessitated a re-examination of the Haldanian concept of decompression which in turn has led to the development of diverse models of inert gas transport for the design of safe and effective decompression schedules (37, 38, 42, 54, 67, 78). In general, the key questions in this area which have not been settled by unequivocal experimental evidence are: (Fig. 13)

(A) Whether the rate of transport of inert gases into and out of the tissues of the body is limited by tissue perfusion or by the diffusion of gases into and out of the intercapillary space (39).

(B) Similarly, gas exchange compartments are generally considered as discrete and parallel units characterized by specific time constants of gas exchange (65, 68, 77) but other interpretations of this concept (40) have also been advanced.

(C) Finally, it has not yet been resolved whether significant inert gas supersaturation can be maintained in the tissues (12, 56) or whether a gas phase separation occurs whenever the dissolved inert gas pressure exceeds the total pressure acting on the body (18, 40).

The majority of diving research laboratories employ perfusion-limited, multiple compartment models of inert gas transport.

Such models permit the relatively straightforward computation of inert gas tissue tensions during and following a particular dive (72). Using the symbols defined in Figure 14, the relationship between alveolar and tissue inert gas pressure is given by a simple differential equation (Fig. 15) which can be integrated and specifically solved for the case where the diver's ascent is either step-wise ($dP/dt = 0$) or linear ($dP/dt = c$). An examination of the perfusion-limited gas transport model (68) shows that the specific time constant of gas exchange of a particular tissue represents the product of its perfusion and the relative distribution of the inert gas between this tissue and blood. The concept of tissue or tissue compartment in this model is therefore not an anatomical but a functional one in that a given specific time constant k characterizes the totality of all combinations of perfusion rate and tissue fat content (which affects the ratio α blood/α tissue), which, on multiplication with one another, yield a value equal to k.

Principal Differences in the Theoretical Treatment of
Inert Gas Transport and of the Formation of Tissue Gas Bubbles

Concept	Interpretation
Rate Limitation of Inert Gas Transport	1. By perfusion of the tissue 2. By diffusion of the inert gas
Theoretical Tissue	1. One of many parallel gas exchange units 2. A single, liquid two-phase system
Appearance of a Tissue Gas Phase	1. On exceeding a metastable maximum pressure of inert gas dissolved in the tissue 2. On decreasing the total pressure to a level below that of the pressure of inert gas dissolved in the tissue

Fig. 13

Explanation of Symbols Used

π = Partial Pressure of Inert Gas Dissolved in the Tissue

t = Time

k = Specific Time Constant of Gas Transport

$t_{\frac{1}{2}}$ = Half-time of Gas Transport = $\ln 2/k$

P = Alveolar Partial Pressure of Inert Gas

c = (constant) Rate of Increase or Decrease of P with Time

Q = Rate of Tissue Perfusion (Volume/Time Unit/Volume)

α = Gas Solubility

Fig. 14

GAS TRANSPORT EQUATION:

$$\frac{d\pi}{dt} = k\,(P-\pi)$$

GENERAL SOLUTION:

$$\pi = e^{-\int k\,dt}\left[\int P e^{\int k\,dt}\,dt\right] + C_1 e^{-\int k\,dt}$$

SPECIAL SOLUTION FOR $\frac{dP}{dt} = c$:

$$\pi = P_0 + c\left(t - \frac{1}{k}\right) - \left(P_0 - \pi_0 - \frac{c}{k}\right) e^{-kt}$$

$$k = \dot{Q}\,\frac{\alpha\ \text{BLOOD}}{\alpha\ \text{TISSUE}}$$

Fig. 15

An interesting property of the solution of the gas transport equation for $dP/dt = c$ is the fact that on continuous ascent (Fig. 16) this equation, somewhat restated, resolves itself into a simple relation between alveolar inert gas partial pressure P, its constant rate of change $dP/dt = c$, the tension of the inert gas dissolved in the tissue, π, and specific-time constant k of the tissue under consideration. If the period of ascent is sufficiently long, the tissue which will determine the maximum safe rate of this ascent is the slowest gas exchange entity in the human body. The maximum permissible difference between P and π for the slowest gas exchange compartment can be determined empirically by an analysis of the decompression requirements of divers who have remained under pressure long enough to completely saturate even their slowest gas exchange compartment with inert gas. Almost invariably it was found that as the saturated diver approaches the surface, his rate of decompression must be reduced to as little as 15 min./ft. or even less, in order to avoid clinical symptoms of decompression sickness.

Inherent in most of these models is the assumption that a finite amount of supersaturation can be tolerated during decompression without provoking formation of gas bubbles. This metastable limit is held to vary with total pressure, tissue half-time (specific time constant) and nature of the inert gas (8, 19, 30, 71). Two proposed limits (76, 77) for slow and fast gas exchange tissues are compared in Figure 17. While Bühlmann's limits for relatively fast tissues are lower than those proposed by Workman, for tissues with half-times between 180 and 240 minutes, Workman's values are far more conservative than Bühlmann's. The late Dr. Hartmann, incidentally, used limiting values similar to those of Bühlmann for slow gas exchange tissues (34).

$$P - \Pi = \frac{c}{k} + \left(P_0 - \Pi_0 - \frac{c}{k}\right) e^{-kt}$$

$$\lim_{t \to \infty} (P - \Pi) = \frac{c}{k}$$

Fig. 16

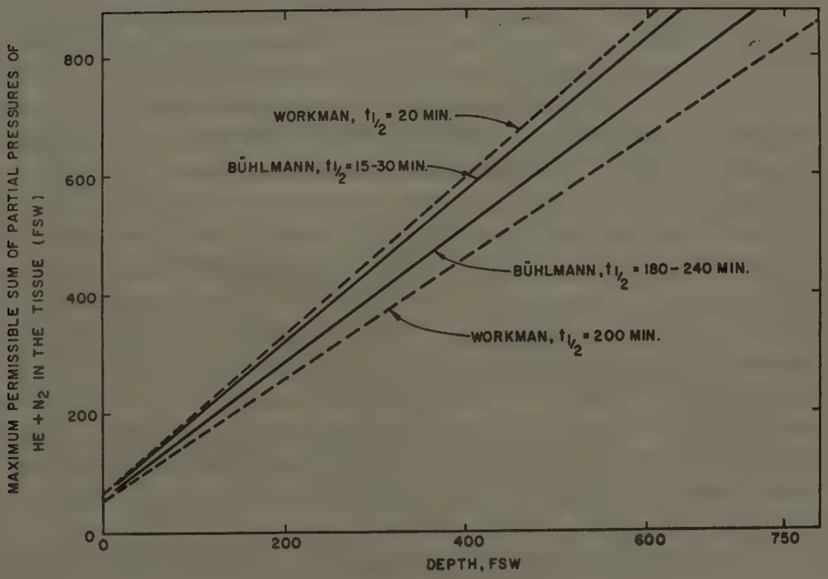


Fig. 17

Viewed in this analytical manner, the choice of decompression regime resolves itself into the selection of the most appropriate method permitting return to surface without exceeding empirically determined or assumed limits of inert gas pressure in the tissues.

In other words, current decompression procedures represent nothing more than the empirical avoidance of clinically significant gas bubbles in the body tissues of the diver being returned to the surface. The processes of inert gas transport, nucleation, bubble formation and growth in body tissues are still poorly understood. To me they seem worthy indeed of the attention of the medical physicist (36, 59, 64).

Saturation diving, even though it is associated with lengthy periods of decompression, represents the most effective use of a diver's time in cases where a considerable amount of work must be performed on a fairly steady basis. Since the period of decompression does not increase with bottom time after the tissues have come to pressure equilibrium with the inspired inert gas, the effective use of the diver's time under water actually increases with increasing bottom time. For relatively short or infrequent dives, step-wise ascent directly to the surface is generally most economical of time. (Fig. 18) In cases where intermittent diving to a particular target depth is required, repetitive excursions from a given saturation depth combine some of the advantages of surface and saturation dives, and contain important elements of both. This saturation-excursion form of diving is relatively new (51) but can be made highly effective by the computer-aided choice of saturation depth in relation to target depth, and of optimized repetitive decompression profiles to saturation depth. In one recent study (31) 24 repetitive excursion decompressions were conducted from 400 to 280 feet after two hours each of heavy work in a wet chamber without incurring symptoms of decompression sickness.

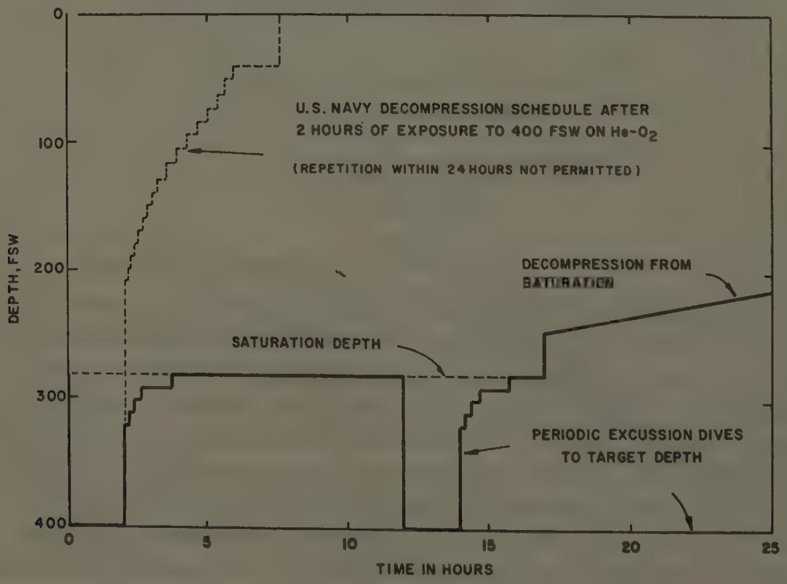


Fig. 18

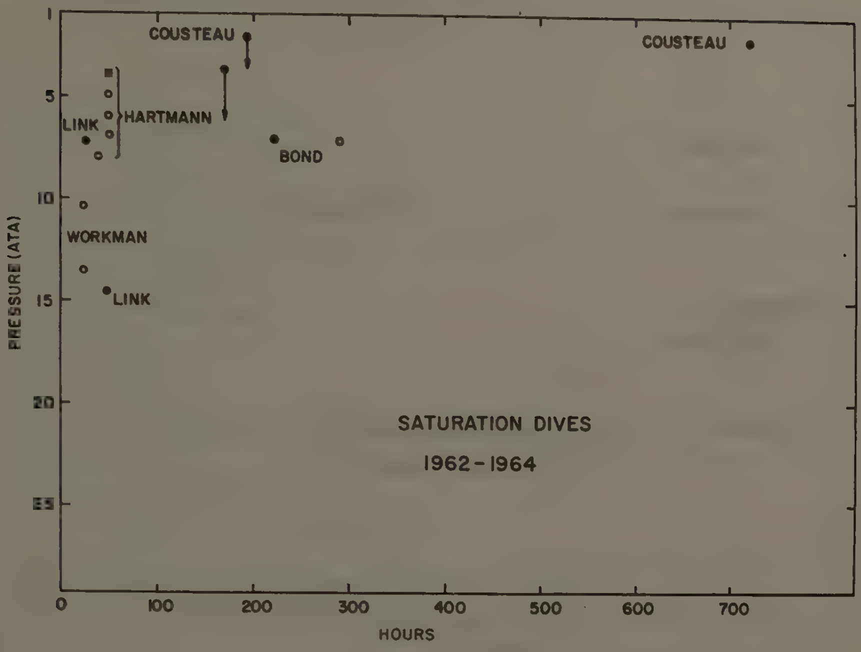


Fig. 19

As formidable as these complexities may appear, there has been rapid and surprisingly orderly progress in probing the limits of man's capability to coexist with the physical facts of deep submergence. This limit can be determined only in extended, carefully controlled chamber and open sea dives. Saturation dives, defined as involving exposures to pressures lasting 24 hours or longer, provide the most appropriate experimental vehicle for such determinations. The need for saturation dives to determine ascent-limiting decompression parameters in man was first pointed out by A. R. Behnke (U.S. Naval Medical Bulletin 35:219, 1937). First attempts to conduct such dives on air at shallow depths in pressure chambers were made during the years following World War II by Behnke, Van der Aue, Duffner, and Workman (personal communication, 1962). Saturation diving became a practical reality only seven years ago (Fig. 19) when Link (25, 52, 74), and Bond (60), and Cousteau (16) conducted their initial experimental saturation dives at sea. The depth of these exposures has increased steadily in the years which followed these pioneering dives (70) (Fig. 20), reaching 650 feet in 1965 (28) (Fig. 21), 702 feet in 1966* (34) (Fig. 22), 825 feet in 1968 (Fig. 23), and 1000 feet earlier this year (Fig. 24). This total body of experience, tabulated as Appendix I to this presentation, indicates that man can survive exposure to reasonably high pressures for extended periods of time without serious detriment. This realization has given rise to a rapidly growing number of subsea habitats at reasonably shallow depths in many countries around the world. Several cur-

---

* This year marks the beginning of Soviet efforts to put men in the sea on a sustained basis (4, 17).

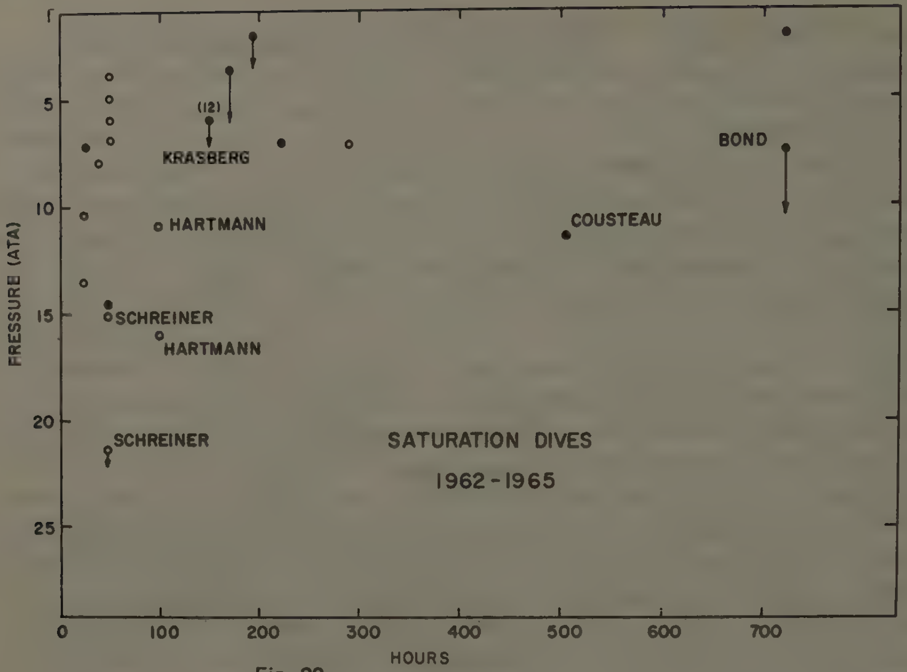


Fig. 20

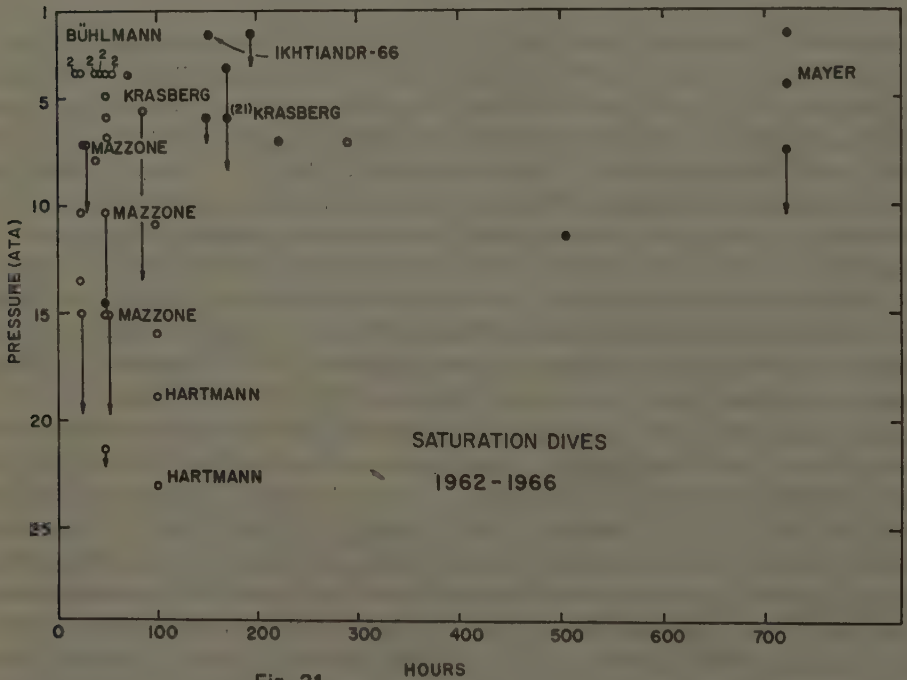
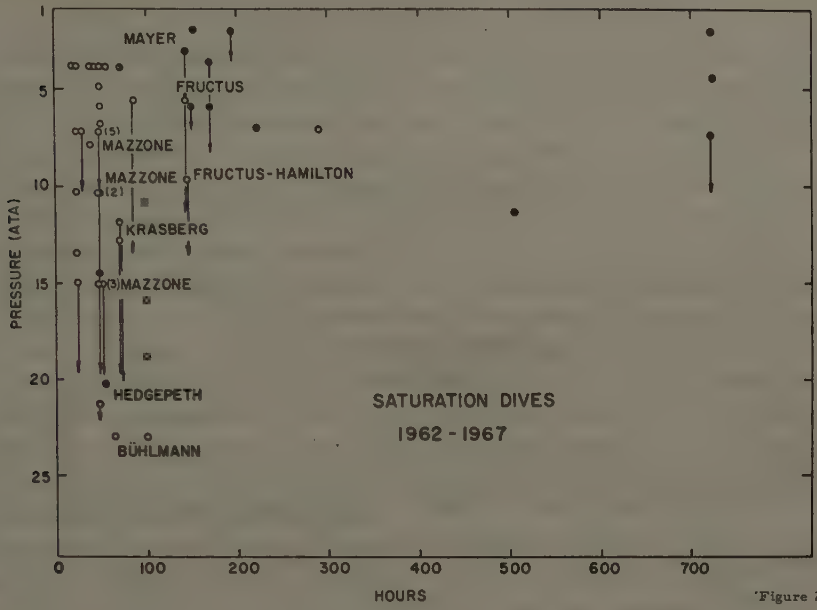


Fig. 21

rent efforts in this area are listed in Figure 25. It is noticeable that the Communist-bloc countries, after a relatively slow start, are now making serious efforts to develop a viable technology in this area. The most ambitious of these



'Figure 22

Fig. 22

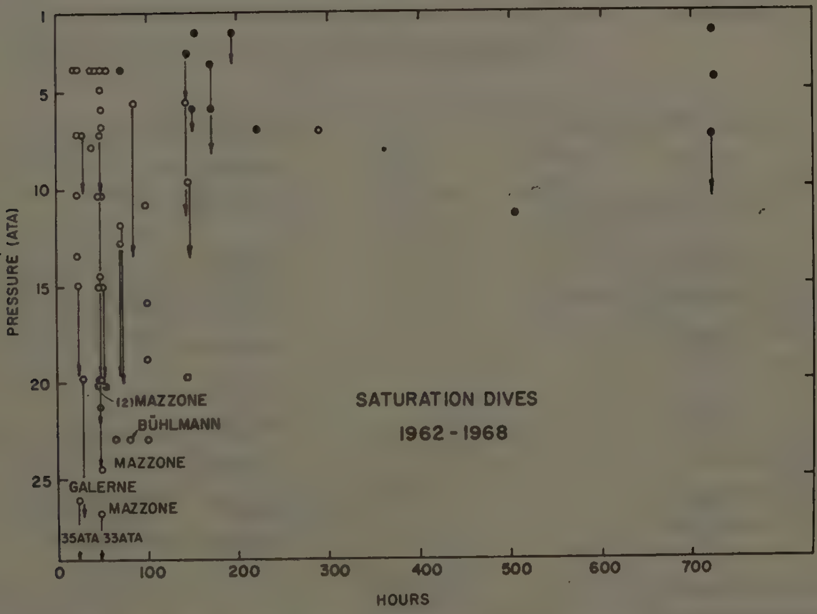


Fig. 23

efforts is the 62-ton submersible laboratory Chernomor, launched last year near Yalta in the Black Sea under the auspices of the Institute of Oceanology of the Soviet Academy of Sciences.

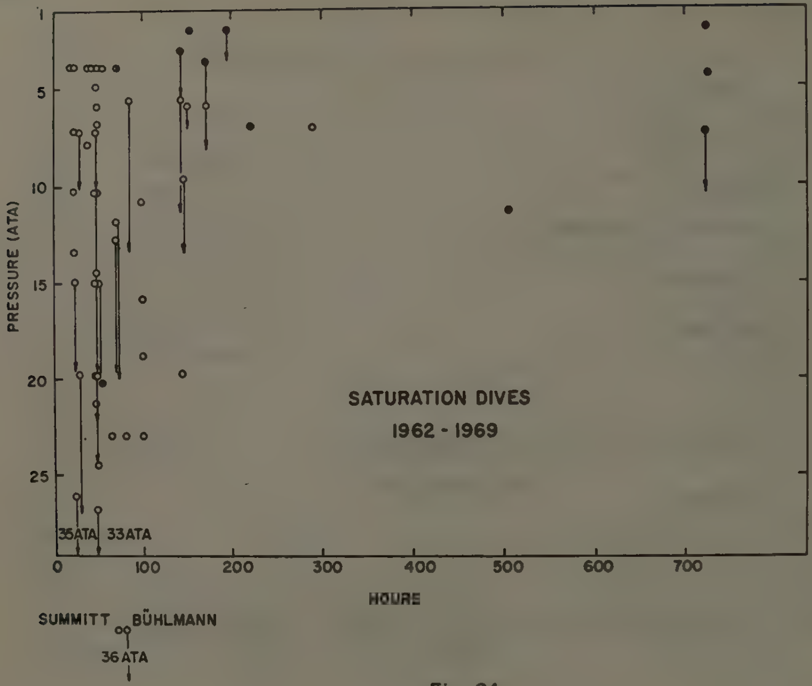


Fig. 24

Examples of Contemporary Manned Underwater Habitats at Shallow Depths

Country	Project
Canada	Sublimnos
Czechoslovakia	Permon
Germany (East)	Malter
Germany (West)	B. A. Helgoland
Great Britain	Bacchus SDM
Poland	Medusa
USA	Tektite
USSR	Sprut Chernomor

Fig. 25

Despite all this progress, the limits of man's ability to perform demanding physical and mental tasks, especially at pressures greater than 20 atm., is only imperfectly understood, and no human being has ever been exposed experimentally to pressures greater than 36 atm. At the highest pressures to which divers have recently become exposed experimentally for short periods of time (Fig. 26), we must begin to be prepared to expect responses of the human body to hydrostatic pressures as such (24). While laboratory animals have survived exposures to pressures equivalent to approximately 5,000 feet (15, 55, 58), it is not at all clear what effect such exposures might have on man. Brauer and his associates (10, 11) describe a "high pressure nervous syndrome" observed in a series of six helium-oxygen chamber dives to pressures of up to 35 atm. This syndrome consists of cerebellar disturbances, somnolence and modifications of the electro-encephalogram. Bühlmann (personal communication, 1969), on the other hand, finds no evidence for such disturbances at this pressure.

High hydrostatic pressures produce marked and easily demonstrable effects in living systems (48, 79). They also increase the equilibrium pressure (22) of dissolved gases which must be kept in mind when considering their activity at great depth. The pressure regions currently being probed by man fall ouside the sphere of interest of the conventional high pressure physicist, yet it seems self-evident that the depth of man's penetration into the ocean will eventually be limited by the effect of pressure *per se* on the human organism. Here, ladies and gentlemen, lies truly a new frontier in medical physics.

In closing I should like to caution you not to lose sight of the fact that mere survival of exposures to ever-increasing pressures or even the return to sea level after such exposures without incident does not constitute progress. We must concentrate to give man, without jeopardizing his safety, the capability of performing in the sea the same wide range of meaningful mental and physical tasks which he is capable of performing on land. Men's intellectual and mechanical skills are, after all, the overriding reasons for his entry into the sea.

Recent Deep Diving Experiments

Year	*Organization*	*Max. Pressure (fsw)*	*Bottom Time (min.)*	*Remarks*
1967	Kantonsspital Zurich	941	120	Stepwise decompression to sea level
1968	U. S. Navy	1025	13	Excursion dive from saturation at 825 feet
1968	Kantonsspital Zurich	957	180	Stepwise decompression to sea level
1968	International Underwater Contractors	1050	5	Excursion dive from saturation at 800 feet
1968	COMEX	1155	15	Stepwise decompression to sea level
1969	Royal Naval Physiological Laboratory	1155	120	Excursion dive from saturation at 1000 feet

Fig. 26

APPENDIX

Chamber and Open Sea Saturation Dives

Year	Saturation Depth (fsw)*	Bottom Time (hrs)	Bottom Gas Mixture			No. of Divers	Project Information			Remarks	Ref.
			% O_2	% N_2	% He		Title	Organization	Director		
1962	200	26	3	—	97	1	Man-In-Sea		E. A. Link	Sea	25
1962	33(→85)	192	21	79	—	2	Précontinent I	M. O. Monaco	J.-Y. Cousteau	Sea	16
1963	36	720	21	79	—	5	Précontinent II	M. O. Monaco	J.-Y. Cousteau	Sea	16
	90(→165)	168	7	—	93	2					
1964	198	288	4	17	79	3	Genesis E	U. S. Navy	G. F. Bond	Chamber	41
1964	300	24	5	10	85	2		U. S. Navy	R. Workman	Chamber	
1964	400	24	4	8	88	2		U. S. Navy	R. Workman	Chamber	
1964	432	49	4	5	91	2	Man-In-Sea		E. A. Link	Sea	52,74
1964	193	222	4	17	79	4	Sealab I	U. S. Navy	G. F. Bond	Sea	60
1964	95	50	13	87	—	2		DVLR	H. Hartmann	Chamber	34
1964	127	50	11	89	—	2		DVLR	H. Hartmann	Chamber	34
1964	159	50	9	91	—	1		DVLR	H. Hartmann	Chamber	34
1964	191	50	8	92	—	1		DVLR	H. Hartmann	Chamber	34
1964	223	40	7	93	—	2		DVLR	H. Hartmann	Chamber	34
1965	318	100	5	27	68	1		DVLR	H. Hartmann	Chamber	35
1965	450	48	2	4	94	2		Ocean Systems	H. R. Schreiner	Chamber	
1965	478	100	3	5	92	1		DVLR	H. Hartmann	Chamber	34
1965	650(→675)	48	2	4	94	2		Ocean Systems	H. R. Schreiner	Chamber	28
1965	205(→300)	360/720	5	17	78	28	Sealab II	U. S. Navy	G. F. Bond	Sea	61
1965	333	504	3	—	97	6	Précontinent III	M. O. Monaco	J.-Y. Cousteau	Sea	
1965	160(→200)	150	7	13	80	11	Smith Mt. Dam	Westinghouse	A. R. Krasberg	Lake 12 saturations	44
1966	200(→300)	24	4	—	96	4	Sealab	U. S. Navy	W. R. Leibold	Chamber	
1966	95	24	21	79	—	2		K. S. Zurich	A. A. Buhlmann	Chamber	14
1966	95	28	21	79	—	2		K. S. Zurich	A. A. Buhlmann	Chamber	14
1966	95	42	21	79	—	2		K. S. Zurich	A. A. Buhlmann	Chamber	14
1966	95	42	21	79	—	2		K. S. Zurich	A. A. Buhlmann	Chamber	14
1966	95	44	21	79	—	2		K. S. Zurich	A. A. Buhlmann	Chamber	14
1966	95	48	21	79	—	2		K. S. Zurich	A. A. Buhlmann	Chamber	14
1966	573	100	3	4	93	1		DVLR	H. Hartmann	Chamber	34
1966	702	100	2	3	95	2		DVLR	H. Hartmann	Chamber	34
1966	36	168	21	79	—	3	Ikhtiandr-66	Donets Club		Sea	4
1966	82	720	21	79	—		Sadko-1	Leningrad Hydromet. Inst.	A. V. Mayer	Sea	17
1966	95	72	20	—	80	2		K. S. Zurich	A. A. Buehlmann	Chamber	14
1966	95	48	20	—	80	2		K. S. Zurich	A. A. Buehlmann	Chamber	14
1966	95	45	20	—	80	2		K. S. Zurich	A. A. Buehlmann	Chamber	14
1966	95	24	20	—	80	2		K. S. Zurich	A. A. Buehlmann	Chamber	14
1966	95	72	20	—	80	4	Capshell	K. S. Zurich	A. A. Buehlmann	Sea	14
1966	150(→400)	84	6	13	81	2		Westinghouse	A. R. Krasberg	Chamber	44
1966	160(→235)	170	6	13	81	23	Cachalot	Westinghouse	A. R. Krasberg	Sea 21 saturations	44
1966	200(→300)	24	4	—	96	4	Sealab	U. S. Navy	W. R. Leibold	Chamber	
1966	300(→450)	48	3	—	97	4	Sealab	U. S. Navy	W. R. Leibold	Chamber	
1966	450(→600)	24	2	—	98	4	Sealab	U. S. Navy	W. R. Leibold	Chamber	
1966	450(→600)	48	2	—	98	4	Sealab	U. S. Navy	W. R. Leibold	Chamber	
1967	200	24	4	—	96	4	Sealab	U. S. Navy	W. R. Leibold	Chamber	
1967	200(→300)	48	4	—	96	5 × 4	Sealab	U. S. Navy	W. F. Mazzone	Chamber	
1967	300(→450)	48	3	—	97	4	Sealab	U. S. Navy	W. F. Mazzone	Chamber	
1967	450	48	2	—	98	2 × 4	Sealab	U. S. Navy	W. F. Mazzone	Chamber	
1967	149(→333)	144	5	15	80	2	Ludion I	Comex	X. Fructus	Chamber	
1967	702	66	3	3	94	2		K. S. Zurich	A. A. Buhlmann	Chamber	76
1967	450(→600)	48	2	—	98	3 × 4	Sealab	U. S. Navy	W. F. Mazzone	Chamber	
1967	702	68	3	3	94	2		K. S. Zurich	A. A. Buhlmann	Chamber	76
1967	380(→600)	72	3	3	94	3		Westinghouse	A. R. Krasberg	Chamber	
1967	350(→612)	72	3	3	94	7	Project 600	Westinghouse	A. R. Krasberg	Sea	
1967	615	54	2	—	98	2	O. S. 600	Ocean Systems	C. H. Hedgepeth	Sea	
1967	82(→147)	144	21	79		2	Sadko-2	Leningrad Hydromet. Inst.	A. V. Mayer	Sea	57
1967	280(→400)	144	3	15	82	2	Ludion II	Comex/OSI	X. Fructus/ R. W. Hamilton	Chamber	31
1968	600	48	1.6	—	98.4	5	Sealab	U. S. Navy	W. F. Mazzone	Chamber	
1968	600	144	1.6	—	98.4	5	Sealab	U. S. Navy	W. F. Mazzone	Chamber	
1968	600(→750)	48	1.6	—	98.4	5	Sealab	U. S. Navy	W. F. Mazzone	Chamber	
1968	600(→750)	48	1.6	—	98.4	5	Sealab	U. S. Navy	W. F. Mazzone	Chamber	
	750	48	0.9	—	99.1	5					
1968	600(→825)	30	0.9	—	99.1	5	Sealab	U. S. Navy	W. F. Mazzone	Chamber	
	825(→1025)	48	0.9	—	99.1	2					
1968	800(→1050)	24				2		IUC	A. Galerne	Chamber	
1968	702	80	1.5	3.5	95.0	3		RNPL	A. A. Bühlmann	Chamber	
1969	1000	77.5	0.9	—	99.1	5	USN-Duke	U. S. Navy	J. Summitt	Chamber	
1969	1000(→1155)	80	1.5	1.6	96.9	3		RNPL	A. A. Bühlmann	Chamber	

---

* 1 fsw = 1/33 atm = 23.033 Torr

REFERENCES

1. Adolfson, J., Deterioration of Mental and Motor Functions in Hyperbaric Air. *Scand. J. Psychol.* **6**, 26, 1965.
2. Agostoni, W., G. Gurtner, G. Torri, and H. Rahn, Respiratory Mechanics During Submersion and Negative-Pressure Breathing. *J. Appl. Physiol.* **21**, 251–258, 1966.
3. Albano, G. *Osservazioni di Fisiologia del Sommazzatore.* Rome: Tipografia Renato Rotolo, 1967.
4. Anonymous, *Seven Days in an Underwater Home.* Sovetskaya Rossiya (USSR), 4 September 1966.
5. Baddeley, A. D., Influence of Depth on the Manual Dexterity of Free Divers: A Comparison Between Open Sea and Pressure Chamber Testing. *J. Appl. Psychol.* **50**, 81, 1966.
6. Beckman, E. L., E. Reeves, and R. Goldman, A Review of Current Concepts and Practices Applicable to the Control of Heat Loss During Water Immersion. *Aerospace Med.* **36**, 136-137, 1965.
7. Bennett, P. B., *The Aetiology of Compressed Air Intoxication and Inert Gas Narcosis.* London: Pergamon, 1966.
8. Bennett, P. B., and A. J. Hayward, Relative Decompression Sickness Hazards in Rats of Neon and Other Inert Gases. *Aerospace Med.* **39**, 301-302, 1968
9. Bennett, P. B., Inert Gas Narcosis. In: *The Physiology and Medicine of Diving and Compressed Air Work.* P. B. Bennett and D. H. Elliott, eds, Baltimore: Williams and Wilkins. 1969.
10. Brauer, R. W., R. O. Way, an dR. Perry, Separation of Anesthetic and Convulsant Effects in Mice Breathing Helium and Hydrogen-Containing Atmospheres at 50 and 150 Atm. *Federation Proc.* **26**, 720, 1967.
11. Brauer, R. W., Electroencephalographic Indices of Central Nervous System Disturbances. In: *Proc. Fourth Symposium on Underwater Physiology,* C. J. Lambertsen, ed., New York: Academic Press, 1970.
12. Brestkin, A. P., Relationship Between the Supersaturation Coefficient of Gas-Liquid Systems and the Tension of Dissolved Gas. In: *The Effect of the Gas Medium and Pressure on Body Functions.* III. M. P. Brestkin, ed., Moscow-Leningrad: Inst. Evol. Physiol. im. I. M. Sechenov, 1964.
13. Bruemmer, J. H., B. B. Brunetti, and H. R. Schreiner, Effects of Helium Group Gases and Nitrous Oxide on HeLa Cells. *J. Cell. Physiol.* **69**, 385-392, 1967.
14. Bühlmann, A. A., P. Frei, and H. Keller, Saturation and Desaturation with Nitrogen and Helium at 4 Atm. *J. Appl. Physiol.* **23**, 458-462, 1967.
15. Chouteau, J., Respiratory Gas Exchange in Animals During Exposure to Extreme Ambient Pressures. In: *Proc. Fourth Symposium on Underwater Physiology,* C. J. Lambertsen, ed., New York: Academic Press, 1970.
16. Cousteau, J.-Y., At Home in the Sea. *National Geographic* **125**, 465-507, 1964.
17. Deryugin, K., Soviet Sealab "Sadko", *Pravda* (USSR), 26 October 1966.
18. Dieter, K. H., On a Mathematical Theory of Decompression. IEEE Transactions on Bio-Medical Engineering, BME, **14**, 124-146, 1967.
19. Doebbler, G. F., R. G. Buchheit, and H. R. Schreiner, *Effect of Helium, Neon, Nitrogen, and Argon on the Relative Susceptibility of Animals to Depth Decompression Sickness.* Washington, 38th Annual Meeting, Aerospace Medical Association, 1967.
20. Doebbler, G. F., M. E. Mohlke, J. M. McMullan, and H. R. Schreiner, Inert Gas Interactions and Effects on Enzymatically Active Proteins. *Federation Proc.* **26**, 650, 1967.
21. Doebbler, G. F., J. H. Bruemmer, and H. R. Schreiner, Influences of High Pressures of Inert Gases Upon Cell Activity. In: *Proc. Third Symposium on Underwater Physiology,* C. J. Lambertsen, ed., Baltimore: Williams and Wilkins, 1967.
22. Enns, T., P. F. Scholander, and E. D. Bradstreet, Effect of Hydrostatic Pressure on Gases Dissolved in Water. *J. Phys. Chem.* **69**, 389-391, 1965.
23. Fenn, W. O., Interactions of Oxygen and Inert Gases in Drosophila. *Resp. Physiol.* **3**, 117-129, 1967.
24. Fenn, W. O., The Physiological Effects of Hydrostatic Pressures. In: *The Physiology and Medicine of Diving and Compressed Air Work,* P. B. Bennett and D. H. Elliott, eds., Baltimore: Williams and Wilkins, 1969.

25. Godley, J., The Long, Deep Dive. *National Geographic* **123,** 718-731, 1963.
26. Golden, R. M., Improving Naturalness and Intelligibility of Helium-Oxygen Speech, Using Vocoder Techniques. *J. Acoust. Soc. Am.* **40,** 621-627, 1966.
27. Hamilton, R. W., Jr., G. F. Doebbler, C. H. Nuermberger, and H. R. Schreiner, Biological Effects of Possible Spacecraft Atmospheres: Exposure of Rabbits to Helium, Neon, Nitrogen, Argon, and Oxygen. *Physiologist* **9,** 197, 1966.
28. Hamilton, R. W., Jr., J. B. MacInnis, A. D. Noble, and H. R. Schreiner, Saturation Diving at 650 Feet. Technical Memorandum B-411, Tonawanda, New York: Ocean Systems, Inc., 15 March 1966.
29. Hamilton, R. W., Jr., Physiological Responses at Rest and in Exercise During Saturation at 20 Atm. of He-O_2. In: *Proc. Third Symposium on Underwater Physiology,* C. J. Lambertsen, ed., Baltimore: Williams and Wilkins, 1967.
30. Hamilton, R. W., Jr., and H. R. Schreiner, Effect of Helium, Neon, Nitrogen, and Argon on the Relative Susceptibility of Animals to Altitude Decompression Sickness. Washington, 38th Annual Meeting, Aerospace Medical Association, 1967.
31. Hamilton, R. W., Jr., P. Fructus, and X. R. Frutus, Physiological Surveillance and Performance Tests During a Seven-Day Exposure to 9.5 Atmospheres Helium-Oxygen with Exercise Excursions at 13 Atmospheres. *Proc. 24th Intern. Congr. Physiol. Sci.,* Washington, 1968.
32. Hamilton, R. W., Jr., and H. R. Schreiner, Putting and Keeping Man in the Sea. *Chem. Eng.* **75,** 263-270, 1968.
33. Hahnloser, P. B., E. Domanig, E. H. Lanphier, and W. G. Schenk, Jr., Hyperbaric Oxygenation: Alterations in Cardiac Output and Regional Blood Flow. *J. Thoracic Card. Surg.* **52,** 223-231, 1966.
34. Hartmann, H., and H.-D. Fust, Langzeit-Druckkammerversuche im Bereich von 4 bis 23 ata. *Draeger-Heft* **265,** 1-18, 1966.
35. Hartmann, H., K.-H. Weiner, H. D. Fust, and R. Seifert, Ueber einen Dauerversuch von 100 Studen bei einem Druck von 11 ata. *Int. Z. angew. Physiol.* **22,** 30-44, 1966.
36. Harvey, E. N., Physical Factors in Bubble Formation. In: *Decompression Sickness,* p. 90, Philadelphia-London: Saunders, 1951.
37. Hempleman, H. V., British Decompression Theory and Practice. In: *The Physiology and Medicine of Diving and Compressed Air Work,* P. B. Bennett and D. H. Elliott, eds., Baltimore: Williams and Wilkins, 1969.
38. Hills, B. A., Thermodynamic and Kinetic Approach to Decompression Sickness. Adelaide: Libraries Board South Australia, 1966.
39. Hills, B. A., Diffusion versus Blood Perfusion in Limiting the Rate of Uptake of Inert Non-Polar Gases by Skeletal Rabbit Muscle. *Clin. Sci.* **33,** 67-87, 1967.
40. Hills, B. A., Decompression Sickness: A Study of Cavitation at the Liquid-Liquid Interface. *Aerospace Med.* **38,** 814-817, 1967.
41. Hock, J. R., G. F. Bond, and W. F. Mazzone, Physiological Evaluation of Sealab II. DSSP Office Publication, U. S. Navy Nortronics Division, Northrop Corporation, Anaheim, Calfornia, 1966.
42. Keller, H., and A. A. Bühlmann, Deep Diving and Short Decompression by Breathing Mixed Gases. *J. Appl. Physiol.* **20,** 1267-1270, 1965.
43. Kinney, J. A., and S. M. Luria, Vision and Visibility. In: *Proc. Fourth Symposium on Underwater Physiology,* C. J. Lambertsen, ed., New York: Academic Press, 1970.
44. Krasberg, A. R., The Evolution of Functional Saturation Diving. In: *Proc. Second Symposium on Underwater Soc.,* London, England, 1967.
45. Kylstra, J. A., I. S. Longmuir, and M. Grace, Dysbarism: Osmosis Causes by Dissolved Gas. *Science* **161,** 289, 1968.
46. Kylstra, J. A., The Feasibility of Liquid-Breathing and Artificial Gills. In: *The Physiology and Medicine of Diving and Compressed Air Work.* P. B. Bennett and D. H. Elliott, eds., Baltimore: Williams and Wilkins, 1969.
47. Lambertsen, C. J., Basic Requirements for Improving Diving Depth and Decompression Tolerance. In: *Proc. Third Symposium on Underwater Physiology,* C. J. Lambertsen, ed., Baltimore: Williams and Wilkins, 1967.
48. Landau, J. V., Hydrostatic Effects on Cellular Functions. In: *Proc. Fourth Symposium on Underwater Physiology,* C. J. Lambertsen, ed., New York: Academic Press, 1970.

49. Lanphier, E. H., Interactions of Factors Limiting Performance at High Pressures. In: *Proc. Third Symposium on Underwater Physiology,* C. J. Lambertsen, ed., Baltimore: Williams and Wilkins, 1967.
50. Lanphier, E. H., Pulmonary Function. In: *The Physiology and Medicine of Diving and Compressed Air Work,* P. B. Bennett and D. H. Elliott, eds., Baltimore: Williams and Wilkins, 1969.
51. Larsen, R. T., and W. F. Mazzone, Excursion Diving from Saturation Exposures at Depth. In: *Proc. Third Symposium on Underwater Physiology,* C. J. Lambertsen, ed., Baltimore: Williams and Wilkins, 1967.
52. Link, E. A., Outpost Under the Ocean. *National Geographic* **127,** 530-533, 1965.
53. Lord, G. P., G. F. Bond, and K. E. Schaefer, Breathing Under High Ambient Pressure. *J. Appl. Physiol.* **21,** 1833-1838, 1966.
54. Lundgren, C. E. G., Experimental Decompression Sickness in Goats in Relation to Predicted Tissue Gas Supersaturation. *Försvarsmedicin* 3, Supplement 3, 1967.
55. MacInnis, J. B., J. G. Dickson, and C. J. Lambertsen, Exposure of Mice to Helium-Oxygen Atmosphere at Pressure to 122 Atmospheres. *J. Appl. Physiol.* **22,** 694-698, 1967.
56. Mackay, R. S., Simulation Studies of the Physiological State Under High Pressure. In: *Proc. Sixth Intern. Conf. Med. Electron. Biol. Eng.,* Tokyo, 1965.
57. Mayer, A. V., V. Ye Dzhus, and A. V. Ignat'yev, The Sadko-2 Underwater Laboratory, *Sudostroyeniye* (Shipbuilding), May 1968, pp. 11-14.
58. Miller, K. W., W. D. M. Paton, W. B. Streett, and E. B. Smith, Animals at Very High Pressures of Helium and Neon. *Science* **157,** 97-98, 1967.
59. Müller, K. G., and S. Ruff, Gasblasenbildung im Organismus beim Druckfall. *Int. Z. angew. Physiol.* **20,** 521-544, 1965.
60. O'Neal, H. A., G. F. Bond, R. E. Lanphear, and T. Odum, An Experimental Eleven-Day Undersea Saturation Dive at 193 Feet. Report ACR-108, Office of Naval Research. Washington, D. C., 1965.
61. Pauli, D. C., and G. P. Clapper (eds.), An Experimental 45-Day Undersea Saturation Dive at 205 Feet. Report ACR-124, Office of Naval Research, Washington, D. C., 1967.
62. Raymond, L. W., Physiologic Mechanisms of Maintaining Thermal Balance in High-Pressure Environments. *J. Hydronautics* **1,** 102–107, 1967.
63. Robinson, E. K., B. B. Brunetti, J. H. Bruemmer, and H. R. Schreiner, Inert Gas Effects on Growth of Mammalian Cells in Tissue Culture. *Federation Proc.* **27,** 706, 1968.
64. Ruff, S., Theoretische Untersuchungen zur Entstehung von Gasembolien und Gasblasen im Gewebe. *Der Anaesthetist* **15,** 317-319, 1966.
65. Ruff, S., and K. G. Müller, Theorie der Druckfallbeschwerden und ihre Anwendung auf Tauchtabellen. *Int. Z. angew. Physiol.* **23,** 251-292, 1966.
66. Schreiner, H. R., R. W. Hamilton, Jr., A. D. Noble, L. A. Trovato, and J. B. MacInnis, Effects of Helium and Neon Breathing on Man at 20.7 Atm. Pressure. *Federation Proc.* **25,** 230, 1966.
67. Schreiner, H. R., and P. L. Kelley, Computation Methods for Decompression from Deep Dives. In: *Proc. Third Symposium on Underwater Physiology,* C. J. Lambertsen, ed., Baltimore: Williams and Wilkins, 1967.
68. Schreiner, H. R., Safe Ascent After Deep Dives. *Rev. Subaquat. Physiol. Hyperbar. Med.* (Paris) **1,** 28-37, 1968.
69. Schreiner, H. R., General Biological Effects of the Helium-Xenos Series of Elements. *Federation Proc.* **27,** 872-878, 1968.
70. Schreiner, H. R., Ein kritscher Ueberblick ueber den gegenwaertigen Stand der angewandten Tieftauchphysiologie. *Int. Z. angew. Physiol.* **27,** 76-98, 1969.
71. Schreiner, H. R., Advances in Decompression Research. *J. Occup. Med.* **11,** 229-237, 1969.
72. Schreiner, H. R., and P. .L Kelley, A Pragmatic View of Decompression. In: *Proc. Fourth Symposium on Underwater Physiology,* C. J. Lambertsen, ed., New York: Academic Press, 1970.
73. Sears, D. F., and E. L. Fuller, Volume Changes of Polar and Non-Polar Liquid Hydrocarbons Exposed to Pressures of Gases. *Resp. Physiol.* **5,** 175-186, 1968.
74. Sténuit, R. P., The Deepest Days. *National Geographic* **127,** 534-547, 1965.

75. Stratton, J., Our Nation and the Sea. Report of the Commission on Marine Science and Engineering. Washington: Executive Office of the President, 1969.
76. Waldvogel, W., and A. A. Bühlmann, Man's Reaction to Long-Lasting Overpressure Exposure. Examination of the Saturated Organism at a Helium Pressure of 21-22 Atm. *Helv. Med. Acta 34*, 130-150, 1968.
77. Workman, R. D., Calculation of Decompression Schedules for Nitrogen-Oxygen and Helium-Oxygen Dives. Research Report 6-65. Washington: U. S. Navy Experimental Diving Unit, 1965.
78. Workman, R. D., American Decompression Theory and Practice. In: *The Physiology and Medicine of Diving and Compressed Air Work,* P. B. Bennett and D. H. Elliott, eds., Baltimore: Williams and Wilkins, 1969.
79. Zimmerman, A. M., Cell Division in Tetrahymena pyriformis: A High Pressure Study. *Proc. Canad. Fed. Biol. Soc.* **10,** 81, 1967.

MAN'S ROLE IN INNER SPACE

George F. Bond

The opinions and assertions contained herein are the private ones of the author and are not to be construed as official or reflecting the views of the Navy Department or the naval service at large.

INTRODUCTION

The exploration and development of the ocean depths, or inner space, deserve equal attention to that given to outer space if only because of the immediate and potential material rewards which will certainly result from intelligent exploitation.

The basic purpose of the U.S. Navy Man-in-the-Sea Program might be simply stated as a broad and continuous effort to extend man's ability as a free agent to explore, exploit, and maintain prolonged existence on the ocean floor to the outer depth limits of the continental shelves, which make up about 14% of the submerged world's surface—a land mass equivalent to the continent of Africa. This area could provide an estimated 85% of the edible protein, petroleum, and mineral requirements of future generations. Additional benefits resulting from manned undersea activities include salvage, ocean floor construction, and even personnel rescue from disabled submarines.

It is the purpose of this essay to describe the potentials of undersea development, the current progress in this field, and the biophysical problems involved. First, the common problems and clear disparities between the programs of inner and outer space are compared. Next, some vital medical factors are discussed. Third, the bioengineering problems which demand the attention of the biophysical investigator, design engineer and fabricator are treated. Finally consideration will be given to future applications.

Bioengineering for Inner Space

In his efforts to gain dominion over the seas of this planet, man has historically sought the means of deeper and longer ocean penetrations incorporating the capability of free-ranging activity. His lack of success in this endeavor is underscored by the fact that in this year two astronauts were able to emerge from a landing capsule into the near-vacuum of space more than a quarter million miles beyond the earth and subsequently walk on the moon. In this same year, the free-diving aquanaut could descend only about 200 meters beneath the surface of the ocean.

The reasons for this disparate state of affairs become obvious upon close examination. Considering the weight of sea water, a pressure differential will be imposed on the man diving to a depth of only 33 feet which will exceed that encountered by the astronaut on the surface of the moon. Relative to decompression problems, it appears that through proper denitrogenation the astronaut can be promised a high degree of protection against bends in all save the most severe accidental circumstances, e.g., rupture of his space suit during extra-vehicular activity. Such is not the case with the aquanaut who is exposed to high ambient pressures of inert gases over considerable periods of time. For this man there is virtually no sure protection against bends; and the cure of this serious disorder does not consist of simply returning to atmospheric pressures of sea level; instead, treatment requires application of added pressure, often many times that of his original exposure.

The astronaut and the aquanaut share some similar problems in addition to the major hazard of decompression sickness. It should be remembered that the astronaut and the undersea swimmer are almost equally weightless, hence subject to Newton's Third Law of Motion; and although the latter is working in a medium about 850 times denser than that of the astronaut on extra-vehicular activity, the explorer of outer space must exert relatively greater effort to overcome the resistance of an inflated space suit.

Again, with both extremes of ambient pressure exposure, a critical feature is that of the inspired breathing mixture, its monitoring and control. The astronaut has a dangerously simple single-gas oxygen supply, with attendant severe fire hazard, whereas the aquanaut requires a multiple-gas mixture, synthesized in proportion to pressure and often with extremely low percentages of the oxygen component, but one in which combustion is not a hazard. Although these differences obtain, control of the breathing mixture remains a common denominator betwixt the two space programs.

Finally it is easy to identify the common problems which exist with respect to crew selection and training, small group interaction, captive atmosphere contaminant control, sensory deprivation and/or alteration, and stress phenomena. However the psychological aspects of undersea exploration will not be considered here.

Addressing specifically the bioengineering problems of undersea development, the question arises: Why, until so recently, has man made virtually no progress as a free agent in the ocean depths—or, more specifically, on the continental shelves of the world? The answer is that standard diving practices call for the undersea worker to depart from sea level, descend to his working depth, and then to return to the surface, the total sequence constituting a single working dive. Such dives, while satisfactory for brief bottom stays with short periods of work, are quite inadequate for the purpose of exploration or exploitation of the continental shelves.

The physical factors which so severely limit the utility of this type of diving stem largely from the established physical laws of Boyle, Dalton, and Henry. Since the amount of inert gas dissolved in the blood and tissue fluids of the diver has a direct relation to the partial pressure of the breathing gas and to the duration of exposure at depth, the deep-sea diver will rapidly accumulate a dissolved gas load in his body, proportional to the depth and time of the dive, which will call for equivalent and often intolerable increases in decompression times for safe return to sea level. The ratio of required hours of decompression to useful working time on the ocean bottom increases with depth in a nearly exponential rather than linear manner. Thus, a single 200-meter dive to the outer margin of a given continental shelf, with only *one minute* of bottom stay could require more than twelve hours of staged decompression for the diver, who would accomplish virtually no useful work during his 60-second visit. From this and similar established schedules it might logically be concluded that working dives below 60 meters are rarely worthwhile in terms of useful human endeavor.

Faced with this situation, diving physiologists sought an oblique approach to the problem. In 1957, the concept of "saturation" diving was propounded. In essence, the saturation diving system postulated that the undersea worker should be provided with a seafloor house, pressurized to ambient sea-water pressure with a suitable breathing gas mixture. From this habitat, the divers could exist directly into the surrounding ocean environment, doing many hours of useful work at this depth, and returning to the underwater house as required. Since no essential change of pressure should be involved, there would be no decompression penalty for these excursions. Instead, the decompression of the "saturated" diver would be accomplished in a single linear ascent to the surface, following days, weeks, or even months of useful work on the ocean bottom. This concept, if workable, seemed to offer the needed breakthrough in manned undersea ventures.

Following the guidelines of the saturation diving concept, animal and subsequently human experimental work was pursued. Within six years it became evident that proper gas mixtures could be provided and adequate decompression schedules promulgated, to permit prolonged human exposures at considerable depths on the ocean floor. Shortly thereafter, volunteer U.S. Navy subjects participated in Operations SEALAB I and SEALAB II, demonstrating that, within the state of present physiological and engineering art, man could become a useful inhabitant of inner space.

In the remainder of this presentation, the following will be surveyed: (1) the external and internal ocean floor environment, including the habitat; (2) the aquanaut himself; and (3) the aquanaut equipment. Biomedical aspects of the work will be treated subsequently.

The External Environment

The ocean environment, or more specifically that of the waters covering the continental shelves of this globe, is clearly hostile. The threatening aspects of this undersea world relate to the water temperature, opacity and to the dangers of the marine biota customarily found in these waters. Each of these hazards will have individual or synergistic effects on the aquanaut and/or his equipment.

As might be anticipated from the unique nature of our program, much of the equipment required for undersea existence was not originally designed for other than sea level, dry environment reliability. When such hardware components are exposed to an environment which is chemically corrosive, highly pressurized, 850 times denser than air, and populated by marine life with fouling and other destructive characteristics, severe engineering problems must be anticipated.

A second problem is posed by the fact that, to date, all attempts at prolonged human habitation of the sea floor have required extensive support from surface vessels, shore facilities, or a combination of both. Thus, all such experiments are subject to the vagaries and hazards of weather and sea state. In consequence an experiment designed to demonstrate man's independence of constraint of the air-sea interface is presently totally dependent on meteorologic conditions of that upper environment.

The Internal Environment

A program to give man a living and working capability to the outer reaches of the continental shelf, about three times the depth of SEALAB II, which rested at only about 200 feet, introduces formidable problems for the physiologist and engineer alike. The physiological unknowns, while still considerable, have been under investigation for many decades, and show promise of resolution in the near future. With regard to engineering problems, however, the situation is less promising. Until very recently there was no incentive to examine the engineering problems associated with prolonged exercises in hyperbaric synthetic gas atmospheres. Equipments available for tests and manned exposures to high pressures were limited to a handful of U.S. Navy recompression chambers which were designed for conventional diving efforts and offered no real inducement to engineering development in this area.

In the course of the early SEALAB experiments, it was at first possible to rely on modified shelf items of hardware to meet the engineering requirements of the program. As greater depths were sought, such makeshift arrangements were no longer possible. Proceeding to fifteen or many more atmospheres of ambient living pressure, a considerable program of engineering development had to be anticipated.

The first problem to be encountered was the increased ambient pressure required for saturated existence at great depths. Here, one had to be assured that all hollow vessels were constructed to tolerate the pressure differentials involved. This was one of the simpler problems, involving only such items as light bulbs, oscilloscopes, and bourdon tubes. Next the inevitable compression of all cellular insulating material, with resulting loss of efficiency, had to be taken into account. Finally, it was appreciated that compression of the atmosphere results in an increase in molecular availability, which in turn greatly augments the factor of heat transfer.

The major problem posed by the synthetic atmosphere relates to the use of helium as the inert gas component of the atmosphere. Virtually no shelf equipment

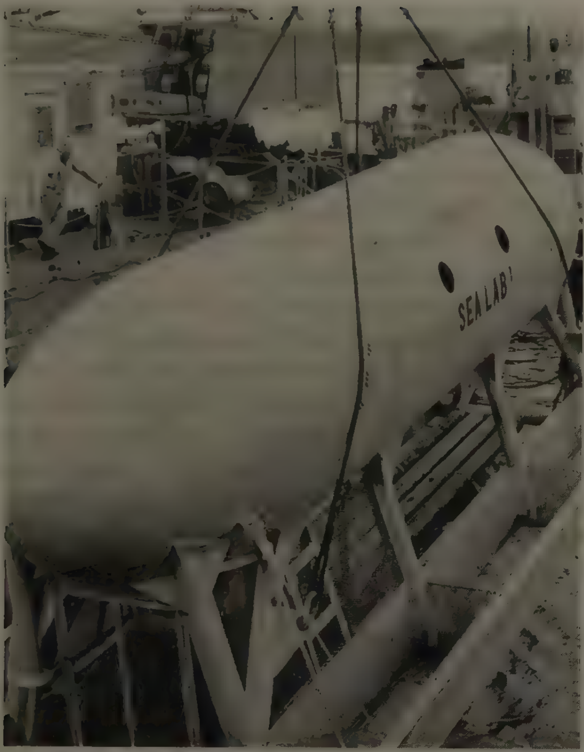


Fig. 1. The SEALAB I habitat being lowered into the water alongside pier in Bermuda, July, 1964. (Official Photograph—U.S. Navy)

—whether mechanical or electronic—will function reliably in a helium atmosphere. Since practically all of the common items used in undersea living are designed to function at sea level pressures, and in an air atmosphere, it is clear that extensive modification or—more often—developmental enginering will be required.

It is necessary to bear in mind the enigmatic and frustrating properties of helium gas. At sea level, it exhibits a thermal conductivity equal to six times that of air; and, at increasing depths, with resultant mounting molecular availability, this problem tends to escalate.

Because of its Reynold's number, so far to the right with respect to our more conventional inert gases, all engineering systems dependent upon turbulent versus laminar flow must be re-examined; and in the case of gas-flow appliances the equation becomes more complex, since gas density must also be considered.

Another problem, not amenable to engineering development, involves the human voice, which is so distorted by helium as to require intervention of an electronic helium voice unscrambler. At the present time, the only workable helium unscramblers are equivalent in size to a 21-inch television set, and require delicate adjustment for each incoming voice. These devices serve to translate the voice of the aquanaut in the habitat as he speaks to the topside control. As we proceed to greater depths, and greater voice distortion, inter-personal communications within the habitat will require translation. At some point beyond fourteen atmospheres of pressure, aquanauts within the habitat will require miniaturized unscramblers no larger than a pack of cigarettes, for convenient conversation.

So much for helium-oxygen communications from a dry SEALAB habitat to the surface control unit. The problem of voice communication in the open sea is vastly more difficult. For many physiological, physical, and environmental reasons, it does not seem likely that open-water voice communications between free divers will be achieved in the foreseeable future.

The internal equipment of the future seafloor habitat continues to present engineering problems, all related to the increased ambient pressure, and to the peculiar characteristics of either helium or hydrogen, whichever inert gas diluent is to be used as the major atmospheric component. Every piece of operating equipment without exception will be radically affected by the presence of lighter inert gases and high ambient pressures. Unfortunately, this is an area which has been largely ignored by the engineering fraternity; hence virtually no currently available equipment will work in the SEALABs of the future. Some examples are in order.

Fig. 2. Aerial closeup of SEALAB II habitat in ocean with divers working on deck. (Official Photograph—U.S. Navy)

The Habitat

In general, the design of the habitat for operation at our ultimate proposed depth should offer no severe problems. Planning for this facility, however, must take into account the fact that autonomy of operation will be required. To this end, a power source must be located on the ocean floor, and engineered to meet the requirements of a fully functional undersea habitat. Obvious candidates for such a power source are modified nuclear reactor-turbine combinations, radioisotope sources, and fuel cells. Each of these systems has points of attraction, and obvious areas of weakness. A well-planned engineering tradeoff study should help to identify the power source of greatest promise. It should be noted, however, that a hydrogen-oxygen cryogenic system may best fit the needs of a fixed habitat. This concept of life support would involve two cryogenically stowed gases operating in a system to offer power via a fuel cell, yield a good breathing mixture, power a catalytic burner to burn atmospheric contaminants, and provide the heat requirements for the habitat. For a mobile SEALAB, nuclear power is the only answer.

Control Systems

If we consider the needs of the man in the SEALAB habitat, it is necessary to provide the temperature-humidity profile best fitted to assure comfortable living conditions in the underwater house. To date, these parameters have not been determined. Systems for control of the atmospheric components are likewise poorly developed. As greater depths are attained, increasingly elegant control of the atmosphere of the habitat will be required. If we wish to maintain the CO_2 partial pressure of not more than 10mm. Hg., it will be necessary to develop more reliable sensor-control systems to meet the greater depth requirements of the program. The engineering situation is further complicated by the problem of increased gas density and associated factors of heat transfer.

At this point the engineering requirements for the ultimate fixed undersea habitat may be projected. The dwelling must be capable of withstanding an internal pressure of 25 atmospheres; it must be designed to support a synthetic atmosphere of either helium-oxygen or hydrogen-oxygen; finally, an autonomous ocean floor power source must be provided.

Looking ahead, some thought must be directed to the development of a mobile undersea habitation, since full utilization of the saturated diver concept will require mobility, for access to any continental shelf of the world. Obviously, the ideal vehicle for this venture would be a nuclear-powered submarine, configured in such a fashion that a large compartment of the vessel's pressure hull would be capable of independent pressurization to ambient sea depth, utilizing a synthetic breathing mixture. This compartment would be occupied by aquanauts, with the capability of departing from the bottomed submarine to do many days of work, whether it be salvage or other engineering/scientific tasks. In such a case, power supply would be no problem, nor would bottom stay be limited. In addition, the factor of weather-proof undersea operations should lend considerable importance to the development. At the present time, it seems most desirable to examine the tradeoffs between new construction and modification of an existing nuclear-powered submarine, with careful consideration of the probable assigned missions of such a mobile SEALAB unit.

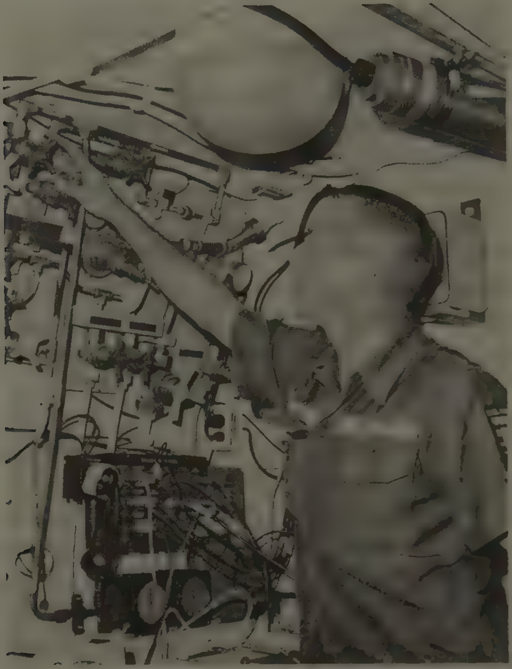


Fig. 3. Life support used by aquanauts in SEALAB II habitat.
(Official Photograph—U.S. Navy)

Aquanaut Equipment

Discussion of aquanaut equipment should touch only on a few select areas which are critical to the undersea program and will require considerable developmental effort. These areas include: protective clothing, communications, visibility aids, work tools, SCUBA equipment, propulsion units, and navigational aids.

Communications continue to be the most unsatisfactory of all systems in use for aquanauts. Communications from the habitat to surface require inter-position of an electronic helium speech "unscrambler" to correct in part the distortions and ablations caused by the helium-oxygen atmosphere. Communications from the habitat to the outside swimmer are possible, for limited range, and at the cost of considerable power. Voice communication from diver to habitat, diver to diver, or diver to surface, however, cannot presently be achieved with available equipment. The problems here involved are formidable: the diver's mouthpiece impedes articulation; exhalation noises provide interference; helium speech distortion precludes intelligible conversation; and omnidirectional broadcasting poses a heavy power requirement. It is quite possible that conventional modes of underwater voice communication must ultimately be abandoned.

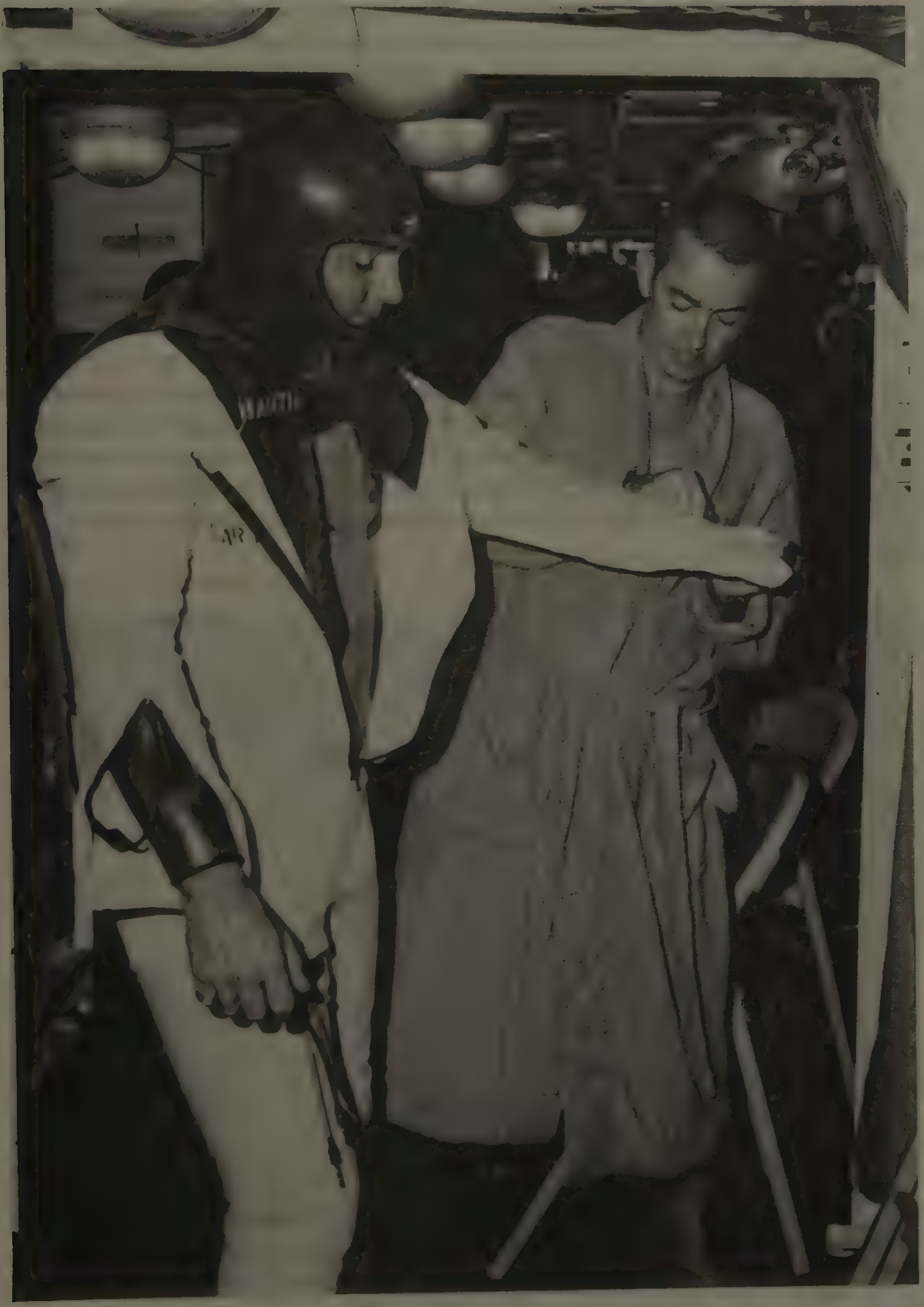


Fig. 4. An electrically-heated wet suit being adjusted in SEALAB II. The suits are heated by battery or by A.C. through the umbilical cord.

(Official Photograph—U.S. Navy)

In the cold environment of continental shelf waters, a major problem of the aquanaut is that of body heat loss. Here, the expected depletion to the water surround is magnified by the thermal conductivity of a helium breathing mixture. At the present time, thermal protection is provided by conventional foam rubber wet suits, which are virtually useless under conditions of SEALAB operations, since the gas cells of the suits become filled with compressed helium and ambient water temperatures are 45°F. or lower. Currently, development is underway with electrically-heated suits which can be powered by batteries or via an electrical cable, in the case of a tethered diver. Additional investigation now includes fluid heating with tube-lined underwear, possibly utilizing a shielded radioisotope heat source.

The problems of decreased visibility in undersea situations deserve considerable attention. In view of the fact that in most continental shelf areas at depths of our contemplated operations, visibility may range from a few inches to 20 feet for unaided vision, it is certain that portable and fixed sources of illumination will be required. These sources should be compact, of long life, and capable of producing as much as 5000 watts. At the present time, the best state-of-the-art has rarely been applied to diver lighting problems. Consideration must be given to pulsed high intensity light, polarized sources, underwater low-luminescent sources, and image intensification.

The underwater worker is faced with many unusual problems, very few of which have demanded solution until quite recently. The conventional hard-hat diver has always used whatever standard suit of tools happened to be at hand. This has been the case for two obvious reasons. In the first place, the tasks assigned to conventional divers have historically been unsophisticated. Secondly, conventional hard-hat suits are so designed that the diver can make himself "heavy," thus remaining fixed in the vicinity of the job, and not easily dislodged. The development of SCUBA gear and the advent of the untethered underwater worker radically changed the situation. With new freedom of motion and vastly increased manual dexterity, the tasks which could be performed by the diver expanded in order of complexity. For these sophisticated technical jobs, an entire new suit of tools required development.

Basically, two factors are to be considered in all future developments relative to aquanaut tools. First, the SCUBA diver-technician is subject to Newton's Third Law of Motion. Therefore, all tools to be developed must have torque-free characteristics if they are rotary in character, or of special design, if they are to be of the impact type. Some progress has been made, but a great deal of work remains to be done.

Very little need be said concerning the diving gear of aquanauts, since this is a familiar subject, and most engineering details are readily available. For aquanauts working at depths of only seven atmospheres, it has been found that two presently available types of diving gear can efficiently be employed. For the free diver, at a depth of 200 feet, the MK VIII and IX semi-closed diving apparatus can be used for about 50 minutes without refill. This SCUBA gear, charged with a predetermined HeO_2 mixture, is so engineered that roughly every third exhalation is dumped from the breathing bag to the surrounding water. Flow rate of the gas to be inhaled is at a pre-set value, determined by a simple formula, in which absolute depth and gas composition are the two critical variables.

For work which can be performed by a tethered diver, a HOOKAH system is most desirable, since it imposes no time limitations on the user. With such a device, the respirable atmosphere of the SEALAB habitat is pumped to the diver in one hose, while his exhaled breath is returned to the habitat, where the CO_2 scrubber system (LiOH or molecular sieve) handles the carbon dioxide accumulation. In the first two SEALAB operations we used this system with reasonable success.

These parameters concern the seven-atmosphere situation. At the 20 to 30 atmosphere depth levels envisioned for the future, the breathing devices described will function only marginally. For the future, a safe, completely closed-circuit apparatus must be developed for the free diver, and new development in compressor-depressor systems will be required for the HOOKAH equipment. To meet the requirement for closed-circuit gear, a completely reliable device for oxygen monitoring and flow control is mandatory. To date, and after more than 15 years of experimental effort, such a device is not available. With respect to the HOOKAH gear, compressor development is not yet adequate to answer the problems posed by the increased density of the gas to be pumped to the diver, who may be required to range several hundred feet deeper than his habitat-placed gas supply systems.

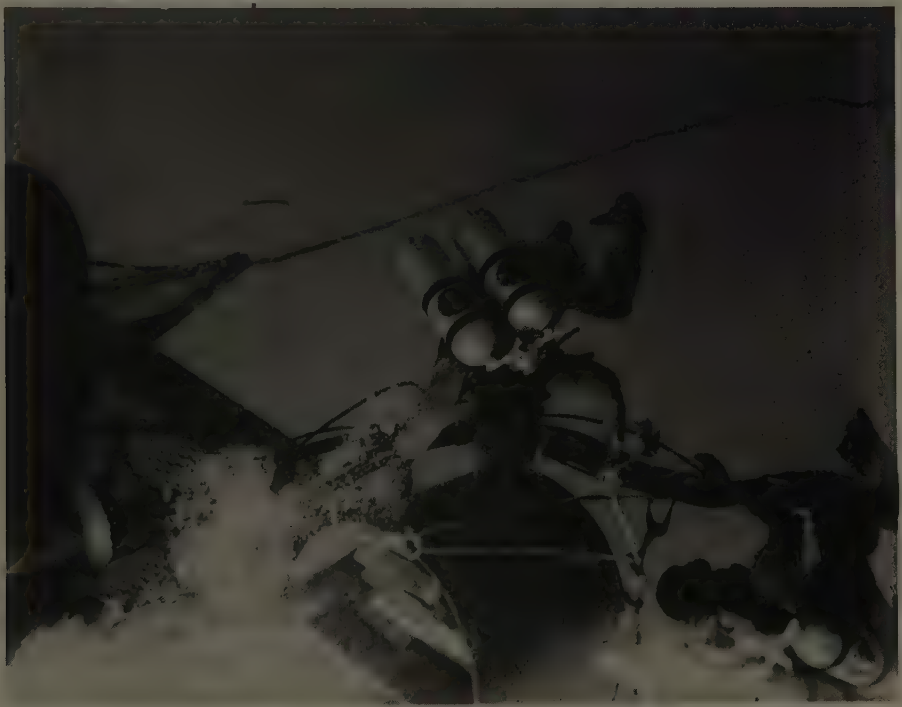


Fig. 5. Divers making final adjustment to the umbilical supply line prior to lowering of SEALAB I off Bermuda.

(Official Photograph—U.S. Navy)

At best, a fully-garbed aquanaut swimmer can only hope to maintain a velocity of about 0.8 knots. Obviously, some devices to provide swimmer propulsion are required. Three sources of locomotion are possible: the "wet diver" propulsion unit; the dry hull small submarine; and a tractor-type of bottom-crawling device. Space does not permit an evaluation of the developmental problems related to these. Suffice it to say that probably all three types of propulsion units will be required for future Man-in-the-Sea efforts.

PHYSIOLOGICAL ASPECTS OF UNDERSEA LIVING

Characteristically, medical research effort in the fields of outer and inner space have common as well as antipodal features. On the one hand, investigation in both areas share problems of synthetic gaseous environment, thermal protection, abnormal degrees of weightlessness, and the psychological problems inherent in autonomous existence in a hostile environment. On the other hand, ambient pressure differences between the two efforts range from vacuum to perhaps 30 atmospheres; work in a medium ranging from near-zero density to one which is 1000 times denser than sea level air; and, in contrast to space capsule living, an underwater existence is one in which man will be deprived to a large extent of senses of sight, touch, smell, and hearing. In this presentation we will treat briefly the environmental problems of undersea living.

The early U.S. Navy experimental effort in the field of high pressure physiology commenced more than a decade ago. This series of experiments, culminating in Project Genesis E*, illuminated and partially solved the problems involved in giving man a capability to explore and exploit the continental shelves of the world. After successful conclusion of this project, the first operational application of the concept was accomplished in Operation SEALAB I, which constituted a firm basis for the current Navy Man-in-the-Sea Program.

In order to pursue the line of experimental effort required to locate man in functional positions on the sea floor, the operational concept of SEALAB I was promulgated as an actual fleet exercise. With approval and assistance of the Secretary of the Navy and Chief of Naval Operations, the first open sea human exercise was commenced in July, 1964.

This project was conducted in approximately 200 feet of water on Plantagenet Bank, 26 miles southwest of Bermuda. For SEALAB I, it was necessary to construct a habitat, approximately 40 feet long and with a 10-foot diameter, which would be largely self-contained, to provide living quarters on the ocean bottom for a group of four men during a proposed stay of about three weeks. This home on the ocean bottom would be supplied with synthetic gas mixture of 4% oxygen, 16% nitrogen and 80% helium, pressurized to the equivalent of the surrounding sea water.

---

* Extreme pressure-chamber studies at the submarine Medical Center, New London, Connecticut which demonstrated that men could live in an artificial helium-oxygen atmosphere at a simulated depth of 200 feet for prolonged periods without harmful effects.

Inside the habitat ambient temperature was maintained at 85° to 88°F, since the thermal conductive properties of helium resulted in a high radiation heat loss from the human occupants. The captive atmosphere was maintained by timely release of oxygen to meet metabolic requirements, and inert gas was added to accommodate for seawater uptake. Carbon dioxide removal was accomplished with a LiOH (lithium hydroxide) scrubber. The habitat was suitably equipped with hot-water shower, standard marine toilet, complete galley facilities, laboratory, and living spaces. Monitoring was accomplished via an umbilical cord which provided communications, electrical energy, closed circuit television, and gas sampling.

Experimental Findings

In the course of the SEALAB I experiment, more than 30 physiological values were derived daily from each subject. These parameters, covering all useful ranges of blood morphology, blood chemistry, basal metabolic functions, body temperature, and general physiological status, gave no evidence of measurable changes in the overall health of any of the underwater subjects. The studies included, in addition to the data recovered from the nine-day undersea exposure, a baseline study of five days of sea level medical data, and an equal span of time after the undersea exposure.

The physiological aspects of project SEALAB I presented no serious problems. Although some new and unexpected facets were uncovered, these were amenable to laboratory study, and presented no real bar to extensions of the SEALAB I concept. Perhaps the most important medical finding of the project was the conviction that, for many years to come, all such undersea manned ventures would require improved coordination of medical and engineering phases of the program, with a great degree of control vested in the medical complement of the operating team. Engineering problems of system design, handling of the complex at sea, and the like, are considerable; but all such problems are interrelated with potential physiological requirements of the experiment.

Two medical findings deserve emphasis: first, the subjects of SEALAB I demonstrated an ability to acclimatize to the temperature losses inherent in HeO_2 living and working in ocean water. Second, there appeared to be a clearly defined slowing of all physiological functions, under conditions of an undersea environment. The first finding is of importance, since evidence from previous experiments promoted the belief that an environmental temperature of 91°F would be required for normal body comfort. In SEALAB I, however, it was found that ambient temperature could be reduced to 85°F after four days of gradual acclimatization. Although this finding did not dismiss the problem of human heat loss, it offered at least a second look at the range of human adaptation and improved design of the submerged habitat.

The atmosphere for the SEALAB I habitation had been preselected at 80% helium, 16% nitrogen, and 4% oxygen. Generally speaking, an attempt was made to adhere to these values. New findings, however, dictated some departures from this set formula. Early in the course of the experiment, a steady drop in nitrogen values was noted, although the phenomenon could not readily be explained.

Helium and oxygen levels, however, remained at anticipated values, with very little makeup required for the former gas, and normal utilization of the latter. In addition, it was noted the requirement for CO_2 removal in SEALAB I was drastically reduced, as compared with previous controlled experiments in pressure chambers. Since basal metabolism of the aquanauts appeared unchanged, it was evident that some new factor was active in CO_2 removal.

As a best possible explanation, we concluded that the small area of sea water exposed in the entrance trunk was serving to absorb considerable quantities of both nitrogen and carbon dioxide. At first glance, this explanation may seem improbable, but certain physical factors must be acknowledged. First, the solubility of CO_2 in sea water is extremely high—about 30 times that of O_2; and secondly, the partial pressures of both CO_2 and N_2 inside SEALAB I as compared with those in the ambient sea water might have been such as to create a favorable gradient for solution of both gases. Helium, being relatively insoluble, would not be taken up in appreciable amounts, which would explain its stability.

This hypothesis must await further laboratory experiments for confirmation; but it would appear that ambient sea water scrubbing might be incorporated in the design of future undersea habitats.

Of physical importance, it was noted that when oxygen levels were held at 4%, or in excess, the aquanauts reported an improved sense of well-being whereas, when levels of 3% to 3.5% obtained, this was not the case. For this reason, although electrocardiographic records showed no hypoxic effects at the lower gradients, oxygen levels in SEALAB I were maintained above 4% (26% effective), for most of the exposure. It seems probable that the increased atmospheric density (approximately 1.6 times greater than sea-level air) may sufficiently impair pulmonary ventilation to require added molecular concentration of oxygen.

A final matter of importance in atmosphere control deserves consideration. In SEALAB I design an air compartment was included, in anticipation of helium speech difficulties, and to provide isolation for the high voltage transformers, which might be potential sources of fire and toxic gases. The compartment, charged with compressed air at the equivalent depth of 186 feet of seawater, was also used for stowage of scuba gear; consequently, the air compartment was occupied on frequent occasions by the aquanauts, who came directly from the helium-oxygen atmosphere of the living compartment.

Within the first 24 hours of exposure, it became apparent that, after denitrogenation in the SEALAB I HeO_2 atmosphere, exposure to compressed air resulted in an immediate and dangerous level of nitrogen narcosis. Briefly, the effect of a visit to the compressed air space was equivalent in narcotic impact to a 350-foot exposure in an air-filled recompression chamber. Once this hazard was recognized, the compressed air was diluted with helium, and the problem resolved. It would thus seem that, once the human body is largely denitrogenated, susceptibility to nitrogen narcosis is significantly increased, and perhaps out of proportion to the involved partial pressures of the gas.

In light of the unhappy experience of Captain Cousteau's group, where a moderately severe anemia was discovered after prolonged exposure to compressed air at only 32 feet of sea water depth, it was gratifying to find no overall hematologic disorders in SEALAB I personnel. Almost certainly, success in this respect can be attributed to the fact that, during the "bottom" phase of the experiment, oxygen partial pressure values were rigidly maintained below 210mm Hg., which may be close to the borderline of prolonged physiological tolerance to this gas.

For the next undersea exercise, entitled SEALAB II, a radically different ocean floor environment was utilized. Seeking a cold, opaque, and extremely hostile surround, we planned our next experiment at a 200 foot depth, on the edge of the Scripps Canyon, off La Jolla, California. For SEALAB II, twenty-eight subjects were carefully trained to occupy, in three successive teams, a newly constructed and vastly more sophisticated habitat. Each team was scheduled for a bottom stay of fifteen days, with two aquanauts prepared for individual stays of thirty days each.

Specific and detailed programs of deep salvage, physical oceanography, seafloor geology, marine biology, and physiology were laid out for each team, and, in addition, excursion dives to the greater depths of the adjacent canyon walls. Environmental conditions were extreme. Water temperature ranged from 54° to 48° F; visibility ranged from one to twenty feet, with artificial illumination; bottom sediments were up to ten feet in thickness; and dangerous marine life was always present. The operation commenced in August 1965.

Although SEALAB II was not designed to yield a large amount of physiological data, it was nevertheless considered prudent to pursue a course of selective monitoring of personnel of each of the three teams. Previous meticuluous physiological studies, conducted through the Genesis series, and subsequently in SEALAB I, had indicated that the great majority of physiological parameters examined would show no significant change under conditions of high pressure and exotic gas mixtures. It was therefore adequate to monitor only those vital functions of our human subjects which might assist topside control in medical management of the experiment.

In consequence, the number of physiological tests performed in SEALAB II was severely curtailed, and the selection of monitored subjects was likewise limited. A program was established to provide for intensive physiological sampling from two or three aquanauts of the first two teams on a daily, alternate daily, then each-third-day basis. Samplings included extraction of about 20 cubic centimeters of blood daily for topside analysis, with appropriate quantities of urine and saliva for additional inspection. In addition, the majority of aquanaut personnel participated in daily studies of pulmonary function, electrocardiographic recordings, body-temperature control, exercise tolerance, and a number of routine physical tests.

By and large, the results of tests on selected subjects in the three teams of SEALAB II were essentially negative. All of the classifical blood chemistries which were performed fell easily within the normal ranges during the experiment. These tests included blood sugars, blood urea nitrogen determinations, creatinines, all serum electrolytes, and calcium and phosphorus values and ratios. Blood morphological values were likewise followed on a regular basis, with no evident deviation from the normal pattern. Urine specimens remained generally within acceptable limits, considering the problems of accurate collection and analysis.

Nevertheless, there were suggestive trends in certain areas which will warrant further intensive investigation under controlled laboratory conditions, and with considerable refinement of techniques. For example, in SEALAB II the red-blood-cell count of the subject (MSC) who was exposed for 30 days of continuous stay to partial pressures in excess of 200mm Hg. of oxygen, showed a linear decrease in red-cell count, although no evidence of cell destruction could be demonstrated. Subsequent studies of a relatively large population, under chamber simulations, demonstrated a consistent drop in red-cell mass for all subjects. Likewise, most of the "stress" enzymes and other indicators of stress were clearly elevated during the first three to five days of hyperbaric exposure.

During SEALAB II, attention was particularly directed to examination of the "stress" enzymes, since these indicators, together with the corticosteroid determinations, had demonstrated greatest lability during past human exposures. These data give provocative evidence of an increased stress effect on the aquanaut subjects during the first three-to-five-day period of undersea exposure, with a slow return to normal values. Of the eight serum enzyme fractions, LDH (lactic dehydrogenase) provoked greatest interest, since values for this enzyme are generally considered to be proportional to a degree of tissue damage. Unfortunately, advanced stratifying techniques could not be utilized with the limited facilities of SEALAB II; hence, the particular organ responsible for the observed LDH overproduction could not be identified. It would appear that "stress" enzyme alterations and possibly serum phospholipids are the most sensitive physiological warning signals available to topside monitors; this fact will be suitably exploited in future undersea experiments.

In summary, three important physiological observations were evident from our studies. First, almost all physiological functions demonstrate a sharp deviation from baseline values during the first three to five days of exposure in all SEALAB experiments. Secondly, the accepted indicators of physiological/psychological stress are generally elevated during the first few days of exposure. Third, a significant decrease in red cell mass has been documented. These provocative observations indicate an urgent need for additional research in this area.

The Future

It is not likely than man will be content to develop a deep-diving capability of only 20 to 30 atmospheres. Without question the commercial and scientific returns of the continental slopes and submarine canyons will demand that we extend man's free-ranging capabilities to 100 atmospheres—perhaps ultimately to the 12,000-foot depths of the abyssal plains, although this greater reach may be of no more than experimental and historical value. For the purpose of this discussion, we will treat the depths of 1,000 to 3,000 feet as practical goals, reserving the greater depths as speculations for narrative consideration.

It seems probable that, within the next three years, we will have developed the necessary technology to perform routine dives to a depth of 1,000 feet. In order to accomplish this goal, a tremendous amount of energy, time, and money must be devoted to engineering and physiological research and development. Yet it can be done, if one may predict with confidence that the breathing mixture at this depth will be approximately 0.9% oxygen, and the balance helium for any prolonged period of existence. Given this controlling factor, the engineering problems and developments can be tailored to reach the desired end result of a system which will function at this level of atmospheric sophistication.

Below 1,000 feet, and jumping to the 3,000-foot level, physiological factors will control. Helium can probably no longer be used as the inert gas constituent, since it may finally be narcotic, and we must move to the left, and bitter end, of the periodic table of elements. Obviously, there remains now only one inert gas which can be used as an atmospheric diluent. This gas is hydrogen; and while Zetterström made significant dives to 14 atmospheres using a hydrogen-oxygen mix two decades ago, we are not yet prepared, from the point of view of physiology or engineering technology, to devise a diving system which would use such a mix. The requirement nevertheless remains, and must ultimately be fulfilled.

In last analysis, we must speculate on man's ultimate free-diving depth capability—an obviously unknown figure. For the purpose of final discussion consider a depth of 12,000 feet, which is close to the average depth of the earth's oceans. It is *possible,* though not necessarily probable, that man may one day swim freely at this depth. It is quite feasible, with proper surgical and engineering techniques, to devise a penetration of the human lung airway through a hole in the windpipe. Through this hole, an oxygenated solution of Ringers solution, suitably buffered to accommodate metabolic CO_2 output, could be pumped through the lungs, supplying the metabolic needs of the deep diver. The system would be engineered to maintain a constant pO_2 of 600mm Hg. in the circulating solution, with the rate of pulmonary circulation appropriately compensated through use of a miniature pump. The foregoing procedures would be performed at four ATA. Next, we flood the remaining head sinus cavities of the aquanaut with any fluid, and put him on an elevator for the deepest of all dives. Theoretically, this diver can walk the abyssal plains of our seas for almost one hour under these conditions, and be returned to the surface with no requirement of decompression, and with only transient, curable damage to his body. This is a way-out projection; but so was saturation diving, ten years ago.

BIBLIOGRAPHY

1. Prolonged Exposure of Animals to Pressurized Normal and Synthetic Atmospheres, R. D. Workman, G. F. Bond, and W. F. Mazzone, Report No. 374, U. S. Naval Medical Research Laboratory, New London, Connecticut, 26 Jan 1962.
2. Prolonged Exposure to High Ambient Pressure, G. F. Bond, PROCEEDINGS of the Second Symposium on Underwater Physiology, ed. C. J. Lambersten, Publication 1181, National Academy of Sciences-National Research Council, Washington, D. C., 1963.
3. Medical Problems of Multiday Saturation Diving in Open Water, G. F. Bond, PROCEEDINGS of the Third Symposium on Underwater Physiology, ed. C. J. Lambertsen, Williams and Wilkins, Baltimore, 1967.
4. Saturation Diving: Man-in-Sea and SEALAB, J. B. MacInnis and G. F. Bond, The Physiology & Medicine of Diving and Compressed Air Work, ed. Bennett and Elliott, Williams and Wilkins, Baltimore, 1969.
5. Man's Future Beneath the Sea, G. F. Bond, Naval Review 1969, ed. F. Uhlig, Jr., U. S. Naval Institute, Annapolis, Maryland, 1969.

COOPERATIVE PROGRAMS AND THE STATUS OF MEDICAL PHYSICS

Robert J. Shalek*, Jack M. Morgan,
Robert Golden, Marilyn Stovall,
and Jackson H. Cundiff

In the United States medical physics has resembled the unassociated and variable practice of medicine. Some medical physics has been good, some has not, and there is much that has been in between. For many physicists so occupied there has been limited communication with other physicists, and also there has been an unclear professional identity; over the last decade the American Association of Physicists in Medicine (AAPM) has greatly helped communications, identity, and the status of medical physicists. Presently there are two federally financed programs which should improve technical communications between physicists and also their professional awareness. The first of these is the Radiological Physics Center (RPC), which is sponsored by the American Association of Physicists in Medicine. It has the task of reviewing the radiation dosimetry on a number of interinstitutional clinical trials in radiation therapy on a national scale. The second program consists of regional medical programs which promise to bring cooperative arrangements within a region, usually the size of a state, in the areas of heart, cancer and stroke. Medical physics usually participates in the cancer programs.

The purpose of the Radiological Physics Center is to assure the correctness of the physical data relating to patients entered into interinstitutional clinical trials involving radiation therapy. There are presently four clinical trials underway and four others in prospect. The purpose of the trials is to test a clinical hypothesis comparing two or more modes of treatment. In a trial, 20 to 30 institutions pool their data in order to derive an early answer. For trials underway there are a total of 101 institutions participating. These are usually major centers which have full-time radiotherapists and physicists. A second purpose of the RPC is to provide feedback to the institutions for the correction of errors which may be uncovered. A third aim is to make instructive findings available generally to the radiological

---

* Supported in part by grant CA 10953 and RM 0 00007 from the National Institutes of Health, U.S.A.

community. To implement these aims a Steering Committee [5] makes policy and the operating personnel [6] carry out the tasks of the RPC. At the onset, acceptable criteria for agreement between an institution and the RPC were chosen as ± 3% on exposure or absorbed dose measurements and ±5% on tumor dose, or the fulfillment of dose prescription. The latter also involves the definition of tumor dose, which can vary from institution to institution.

Earlier surveys of machine calibration or the fidelity of fulfillment of radiation prescription have revealed sizable discrepancies between institutions that were at times real and at times only apparent, being due to differences in definition of tumor dose (2,3,4,5,6). It is probable that in an institution errors in the fulfillment of radiation prescription are soon compensated for by changes in prescription based upon clinical judgment. However, when an institution with a dosimetry difficulty enters a clinical trial and treats according to a protocol, the error may cause mistreatment or affect the conclusions drawn from the trial. Usually persistence in an old error is less hazardous to the patient than an ill considered correction or change in prescription or dosimetry.

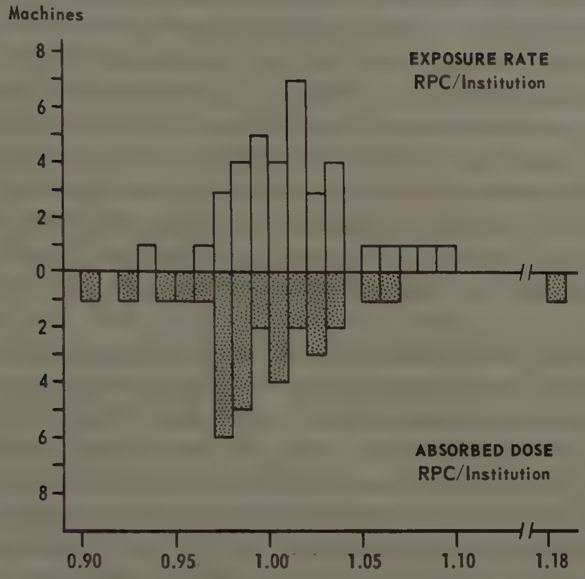
The prime method of institutional review is by means of a visit by an RPC physicist to the institution. The physicist will make measurements, derive information on the technique of treatment, and follow through dose calculations made on cases using processes which the institution itself employs. The second method of review is by the simulation of a radiation treatment cycle utilizing an anatomical phantom loaded with thermoluminescent dosimeters. This method is completely independent of the first, and in a sense redundant, yet may point to deficiencies in the calculative method. However, if both methods indicate the same conclusion, one then has a secure position concerning the conclusion. The third way of deriving information related to an institution is the review of dosimetry calculations for all cases entered into the interinstitutional clinical trial. Here, in addition to the errors in method which may exist, one finds the accidental errors of filling out of forms and of transferring data from hospital charts to report forms.

The findings thus far are shown in Figure 1. There have been 43 institutions which have been visited. Of these, six have been found to fall outside of the ±5% criteria for the fulfillment of dose prescription. These six have entered corrections into their methods; three others are apparently outside the criteria and are in the negotiation stage to seek resolution of the discrepancies. When an error is discovered the RPC indicates the errors, suggests remedies, but recommends that the institution return to its primary sources to develop the corrections themselves. The RPC then again acts as a checking agency to review the corrections. In this way, the institution does not become dependent upon the RPC for its standards. The range of errors has been from −10% to +18%. The most common types of error are listed in Table I. A future publication will consider these in detail.

---

[5] Steering Committee of the Radiological Physics Center: John R. Cameron, Robert O. Gorson (ex-officio), Lawrence H. Lanzl, John S. Laughlin (Chairman), Mary Louise Meurk, Fearghus O'Foghludha, Jacques Ovadia, Robert J. Shalek (ex-officio), Marvin M. D. Williams, Ralph Worsnop.

[6] Operating personnel: Jackson H. Cundiff, Robert Golden, Robert Key, Robert J. Shalek (Director).



Bar graph indicating the number of radiation machines at various institutions and level of agreement with the Radiological Physics Center in exposure rate measurement and tumor dose fulfillment.

Fig. 1

TABLE I

Major Sources of Error

1. Incorrect chamber correction factors.
2. Inappropriate roentgen-rad factors.
3. Differences in irradiation time and time set during calibration.
4. Reliance upon incorrect experimental dose distributions.
5. Non-use of attenuation factor—Cobalt-60.
6. Lack of allowance for fall-off across large fields.
7. Miscalculation of depth dose for oddly shaped fields.

Editorially, from the findings thus far, a message for physicists is appropriate. It is common in large institutions to find that no single person knows all of the steps involved between dose measurement and the time set for treatment of a patient. Not the radiotherapist, the physicist, the dosimetry technician, nor the technician or nurse positioning patients on the machine is fully aware of the whole chain of events. The person most qualified to review all of these steps is probably the physicist. It is a task which the physicist should do on a regular basis—that is, reviewing types of treatment from beginning to end.

The second message is to establish procedures for uncovering arithmetic errors. In institution after institution one finds arithmetic errors in charts. A regular routine checking method is worth the time and expense.

Regional Medical Programs in Physics are distinct from the Radiological Physics Center. These programs are typically statewide and will function in cooperation with other programs, usually in radiation therapy. There are six or seven programs in medical physics now funded, and there are others which have applied for support. The program in Texas was not the first but was one of the early programs, and is one in which the environment is favorable for such a program. That is to say, it is an area in which there are not great rivalries between physicists and yet it is an area where there is a reasonable total number of medical physicists for the job at hand.

In Texas, with a population of 11 million and many square miles, there are 36 medical physicists from 19 institutions, 43 Cobalt teletherapy irradiators, 11 high voltage X-ray sets, 700 therapy X-ray machines and 6,300 diagnostic X-ray machines. The voluntary organization which has been formed by the physicists is called the Texas Regional Medical Physicists. It has a Policy and Planning Committee[7], which is the major governing committee, and operating personnel[8]. A prime point in the formation and philosophy of this organization is that the organization serves the individual physicists and does not serve the medical community directly. Thus it places in the hands of the physicist methods which he as an individual could otherwise not offer; it amplifies his capabilities. It is a voluntary organization, it is not incorporated, and its policies are determined by a vote of all physicists concerned. Thus, it is not a crushing bureaucracy. It is hoped that after two to four years of federal support that the organization will be self-supporting from services rendered and not require grant funds for continuation.

Several functions are being undertaken presently by the Texas Regional Medical Physicists.

Calibration of radiation measuring instruments. The existence of a regional station with carefully maintained radiation standards has the endorsement of the National Bureau of Standards, which endeavors to give rapid and frequent calibration of transfer instruments upon which the regional station depends. Usually the physicists bring their instruments to the meetings of the TRMP and intercomparisons are made during the meeting. In addition to instrument calibration, the use of the instruments is discussed with the user. The user will leave with his instrument and will receive his calibration report in due time.

Mailed verification dosimeters. In use are thermoluminescent dosimeters (TLD) sent by mail for exposure or absorbed dose verification of radiation therapy machines. They provide a check upon the radiation machines as well as the instruments and measurement practices of physicists. One of the major problems for institutions which do not have a regularly employed physicist is the frequency of machine calibration. It is common practice for therapy machines in these circumstances to be calibrated once a year or less. It is possible for x-ray machines, as well as teletherapy irradiators, to give appearance of operating correctly when in fact the radiation dose rate has changed significantly. Mailed dosimeters permit the physician or his technician to irradiate dosimeters in a defined geometry and return the TLD to the center. The results then are sent to the physicist concerned who determines whether a machine has changed since his last full calibration. The TLD is not meant as a calibration method, but simply as a check method to call attention to significant changes in radiation parameters.

---

[7] Policy and Planning Committee: Stewart Bushong, John Domatti, Ralph Griffin, Jack Morgan (Chairman), Robert J. Shalek and Robert G. Waggener.

[8] Operating personnel: Eddie Marie Childers, Karyn Coleman, Robert J. Shalek (Director), Marilyn Stovall.

Interstitial, intracavitary and external beam dosimetry calculations. A network for providing dosimetry calculations for individual patients at different locations utilizes telecopiers. These machines have the capability of transmitting graphical information over an ordinary telephone by sound communication requiring six minutes for an 8½ x 11 page. The originator may send to a central facility the patient contour, copies of radiographs from a radium implant, or coded computer input. He receives back the isodose plan which is drawn by pen and ink by the computer. The service includes treatment plans for cobalt-60 gamma rays and x-rays above 2 Mv. It also includes interstitial and intracavitary radiation treatment. Presently there are 11 telecopiers in Texas. Isodose or decrement line data for individual machines are derived experimentally prior to an institution's entry into the network.

Standardization of procedures. Standardization of procedures which physicists use is derived from committee action and is disseminated in workshops, held twice yearly.

Future activities may include methods for determining the performance characteristics of diagnostic x-ray units, calibration of beta applicators, and availability for loan of specialized apparatus such as anatomical phantoms and isodose measurement instruments.

What are the results of these two types of program likely to be? First of all, it is probable that both of them will improve the average quality of medical physics in the United States. Secondly, these cooperative programs offer physicists the opportunity to demonstrate that they are prepared for professional responsibility. The status of medical physicists in the United States has not been as satisfactory as that of the medical physicists in Great Britain, for instance. In many cases, physicists have been completely dependent upon particular physicians; the term, "my physicist," has emotional connotations for many.

Regional medical programs place the physicists in a favorable position. When changes take place in organized medicine, physicists who have a functioning regional organization will be heard if they know what they wish to say. These regional programs offer the physicists the most immediate challenge and opportunity to improve the status of medical physics. However, in the long run, departments of medical physics or departments of physical science applied to medicine are a sure way to improve the status of the profession and to encourage bright young men into this area of endeavor. Within the next decade, opportunities for these departments will occur. In major institutions there is an increasing depart-mental separation of Diagnostic Radiology, Therapeutic Radiology, and Nuclear Medicine. When this occurs, it makes sense to consolidate the physicists and perhaps engineers and mathematicians into a separate department. There are at least two American institutions that are now in the process of creating departments of medical physics or physical science applied to medicine. These will add to the two others that presently exist. One may predict that there will be opportunities for others in the next few years.

Recently the executive board of the AAPM voted to invest a considerable financial expenditure to assist in the organization of the new Alliance for Engineering in Biology and Medicine. This organization should provide the opportunity for medical physicists to discover bioengineering and vice versa. It is therefore recommended that American medical physicists fully understand their professional situation and aims so that when opportunities for the improvement of the profession occur they will be able to respond. Most immediately, those opportunities are in regional medical programs and more long range in departments of physical science applied to medicine.

Let us, however, maintain perspective and realize that almost every profession regards itself as overworked, underpaid, and underappreciated. Thus, not just medical physics, but probably all professions feel that they are struggling against odds. Aspirations for professional improvement should be accompanied by a willingness to give performance of a higher professional quality. We should expect that progress is not likely to be steady, but will be modulated by the rough and tumble of human affairs.

REFERENCES

1. Ehrlich Margarete, and P. J. Lamperti, Uniformity of High-Energy Electron Beam Calibrations. *Physics in Medicine and Biology* **14,** 305-314, 1969.
2. Holt, G., P. R. Almond, and M. L. Meurk, Intercomparison and Calibration of High Energy X-Rays and Electrons. *Annals of the New York Academy of Sciences* **161,** 133-138, 1969.
3. Lescrenier, C., R. Benbow, A. J. Stacey, and W. G. Gunn, Observations of Treatment Parameters for Megavoltage Generators. *Physics in Medicine and Biology* **12,** 117, 1967.
4. Loevinger, R., P. M. Pfalzner, H. Eisenlohr, S. A. Malo, A. Sanielevici, and J. Nagl, The IAEA Program in Medical Radiation Dosimetry. *Annals of the New York Academy of Sciences* **161,** 158-167, 1969.
5. Spiers, F. W., and W. J. Meredith, Statement of Dosage in Megavoltage Radiation Therapy: Recommendations of the Faculty of Radiologists. *Clinical Radiology* **13,** 163-166, 1962.
6. Worsnop, B. R., Dose Delivered to the Mediastinum; an Intercomparison between Centers. *Radiology* **82,** 1062-1066, 1964.
7. Worsnop, B. R., Phantom TLD Comparison for Cooperative Radiotherapy Trial. *Radiology* **91,** 545-553, 1968.

COMPARTMENTAL MODELING

Mones Berman

INTRODUCTION

Although the present mainstream of biological research is at the biochemical level there is an ever increasing interest in the kinetics of metabolites within complex systems. Isotopic experiments are performed on cell populations, isolated organs and whole animals, and the data are analyzed within the framework of complex models. Although one may question the adequacy and the interpretations of some models, most model builders try through proper modeling techniques to relate their models to physiology and biochemistry. There is also an increasing awareness on the part of the biologist of the important role that models can play in understanding the function of complex, interactive systems. Modeling has not yet reached the stage, however, to serve as a routine tool in the diagnosis and treatment of patients—a role it hopefully will play in the future.

A compartmental system is a system composed of a finite number of discrete, interacting states. Particles within a single state are considered indistinguishable from each other and subject to the same forces and transition opportunities. A schematic representation of such a system is shown in figure 1. The circles represent states and the arrows with solid lines are transition pathways. The

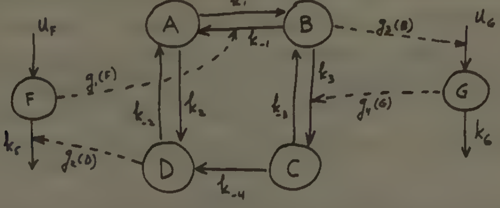


Fig. 1

arrows with dotted lines represent controls that metabolites can have on various transition pathways. The set of differential equations describing the system in figure 1 can be put in the following form:

$$\frac{df_A}{dt} = -k_1 f_A - k_2 f_A + k_{-1} f_B g_1(F) + k_{-2} f_D$$

$$\frac{df_F}{dt} = -k_5 g_2(D)\, F + U_F$$

etc. . .

The functions $g_i(R)$ may be simple or complex depending on the nature of the control mechanisms. k_i are constants of the reaction mechanism, and U_X are input or driving functions.

Let me emphasize the fact that biological systems are, in general, non-compartmental and a compartmental approach is but one approach to modeling. Compartmental models do not necessarily imply corresponding physiology. Several considerations tempt one to conceptualize in terms of compartments:

1) Because of limited resolution in the data the system behaves like a finite number of compartments; 2) Compartmentalization happens to agree with physiological or biochemical conceptualizations that the investigator has about the system; 3) For purposes of analysis knowledge of the correct nature of the system is not necessary and any convenient mathematical formalism that can correctly predict the responses is acceptable; 4) The mathematics of compartmental systems involves ordinary differential equations and is frequently more convenient and easier to use than partial differential or integral equations.

In the final analysis the validity of a model depends on the *purpose* that it serves. For some purposes the type of model assumed is of little consequence provided it fits the data adequately. For other purposes it is most critical.

Biological systems are highly nonlinear, (reactions are not necessarily first order) and they are frequently not in a steady state. A great deal of simplification can be achieved in the experimental design and in the analysis and interpretation of result through linear perturbations, at the expense, however, of information per experiment (1). For example, the description of changes in states due to several non-linear processes may be of the form

$$\frac{dx_1}{dt} = -k_1 x_1^2 + k_2 x_1 x_2 - k_3 \tag{1}$$

The equation contains zeroeth and second order reaction terms. If one labels the components of the system with a tracer, then each x_i can be represented as the sum of two moieties, x_i^- and x_i^* where x_i^- is the stable or tracee (23) material and x_i^* is the tagged material. Setting $x_i = x_i^- + x_i^*$ one gets

$$\frac{d(x_1^- + x_1^*)}{dt} = -k_1(x_1^- + x_1^*)^2 + k_2(x_1^- + x_1^*)\,(x_2^- + x_2^*) - k_3 \tag{2}$$

When x_i^* is very small compared to x_i^* it behaves like a tracer, and the equation for x_i^* may be written as

$$\frac{dx_1^*}{dt} = (-2k_1 x_1^- + k_2 x_2^- - \frac{k_3}{x_1^-} x_1^* + (k_2 x_1^-)\, x_2^* \tag{3}$$

Equation (3) is a linear differential equation in the tracer variables x_i^* since they appear as first order terms only. This demonstrates the well known fact that a gentle perturbation of a nonlinear system behaves linearly. The coefficients in the equation are functions of the tracee (the non-labeled material).

When the system is in a steady state the x_i^- are constant and the equation reduces to a first order differential equation with constant coefficients

$$\frac{dx_1^*}{dt} = -\lambda_{11} x_1^* + \lambda_{12} x_2^* \tag{4}$$

Where $\lambda_{11} = (2k_1 x_1^- - k_2 x_2^- + \frac{k_3}{x_1^-})$ and $\lambda_{12} = (k_2 x_1^-)$

Under these conditions tracer data, if sufficiently complete, lead to the solution of a sub-system having new variables λ_{11} and λ_{12}. The relations between the λ's and k's and the x_i^-, as given in equations (3) and (4), are not apparent from the single tracer experiment. They may be derived from a series of tracer experiments performed at various steady states. Thus computational and modeling simplification are obtained with tracer experiments at the expense of information per experiment.

The bulk of compartmental analysis done in the past dealt with tracer subsystems. There is growing interest, however, in complex non-linear control systems. Since the mathematical tools for dealing with non-linear systems are meager, computers play an important role in the modeling efforts.

COMPARTMENTAL THEORY

For linear systems with time invariant parameters compartmental theory is relatively simple. To apply it, however, is more complex due to the following factors: data are incomplete, data are imprecise and constraints known from the physiology or biochemistry and from the experimental techniques must be imbedded in the analysis.

In matrix notation the set of differential equations describing the system can be written as

$$\dot{x} = -\lambda x + u$$

x is a column vector of responses, λ is a square matrix of rate constants and u is a vector of inputs. λ_{ij} is the rate constant for transitions into compartment i from j with a negative sign. λ_{jj} is the sum of all rate constants representing losses from compartment j.

The solution (4) of the set of differential equations given that u is a vector of impulses, x_o, (initial conditions)—is

$$x = e^{-\lambda t}x_o$$

where $e^{-\lambda t}$ is a square matrix of responses. The i^{th} row and j^{th} column element is the response in the i^{th} compartment to a unit impulse input into the j^{th} compartment.

Expressed in terms of eigenvalues and eigenvectors, for a diagonalizable matrix λ

$$x = Be^{-\alpha t}B^{-1}x_o$$

where B is a matrix of eigenvectors and α is the diagonal matrix of eigenvalues. This can further be reduced to the form

$$[x] = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \ldots \\ A_{21} & A_{22} \ldots \\ \\ \ldots\ldots & \ldots\ldots \\ \\ \end{bmatrix} \begin{bmatrix} e^{-\alpha_1 t} \\ e^{-\alpha_2 t} \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

where A is a matrix of eigenvectors normalized for initial conditions. Thus, the solution can be written as a sum of exponentials,

$$x_i = A_{i1}e^{-\alpha_1 t} + A_{i2}e^{-\alpha_2 t} + \ldots \qquad (i = 1, 2, \ldots$$

Properties of the eigenvalue and eigenvector matrices as a function of the λ matrix have been discussed by Hearon (33).

When the λ matrix is diagonalizable a simple relation exists between the parameters of the system, the λ's and the eigenvalues and eigenvectors (12)

$$\lambda = A\alpha A^{-1} = B\alpha B^{-1}$$

When the λ matrix is not similar to a diagonal matrix, it can be reduced by a similarity transformation to the Jordan Canonical Form **J**:

$$\lambda = PJP^{-1}$$

and the solution may be written as

$$x = e^{-\lambda t}x_o = Pe^{-Jt}P^{-1}x_o$$

which is in the form of sums of terms that are products of exponentials and powers of t:

$$ct^k e^{-\alpha_j t}$$

When the components of the input vector is time dependent, the solution of the set of differential equations

$$\dot{x} = -\lambda x + u(t)$$

is given by (4,54,55)

$$x = e^{-\lambda t}\int e^{\lambda t}u(t)dt + e^{-\lambda t}x_o$$

By proper transformations this can be written as

$$x = \int_0^t e^{-\lambda(t-\tau)}u(\tau)d\tau + e^{-\lambda t}x_o$$

The integral $\int_0^t e^{-\lambda(t-\tau)}u(\tau)d\tau$ is a convolution integral. It is the convolution of $e^{-\lambda t}$ with $u(t)$, frequently written as $[e^{-\lambda t} * u(t)]$. The convolution integral is commutative so that $e^{-\lambda t} * u(t) = u(t) * e^{-\lambda t} = \int_0^t u(t-\tau)e^{-\lambda t}d\tau$.

The general solution for x is thus the superposition of two components: one due to the input function $u(t)$ and the other due to the initial conditions vector x_o. It also expresses the very important convolution concept that given the response $e^{-\lambda t}$ of a system to a set of unit vectors, the response to any input vector $u(t)$ can be predicted. It follows from the general solution that, given that the response in compartment i due to a unit injection into compartment j is $W_{ij}(t)$, the response $X_{ij}(t)$ in the i^{th} compartment due to an arbitrary input $u_j(t)$ into compartment j is also given by the convolution integral

$$X_{ij}(t) = \int_0^t W_{ij}(t-\theta)u(\theta)d\theta = W_{ij}(t) * u(t) = u(t) * W_{ij}(t)$$

W_{ij} is known as a *weighting function.*

A very important special case of the above is $u(t) = u_c =$ constant, i.e., the input vector is a 'constant infusion'. The solution reduces to

$$x = \lambda^{-1}[I - e^{-\lambda t}]u_c + e^{-\lambda t}x_o$$

and as $t \to \infty$ we get the well known steady state solution

$$x(\infty) = \lambda^{-1}u_c$$

where λ^{-1} is the inverse of the λ matrix. When u_c contains unity in the i^{th} row and zeroes everywhere else, the i^{th} column of λ^{-1} is the solution vector. Thus the j^{th} row, i^{th} column element of λ^{-1} is the solution for compartment j due to a unit infusion into compartment i.

More extensive discussion of steady state solutions have been presented by Gurpide et al (31). Multiple input-response systems were also treated by Stephenson (54), and by Perl et al (39).

The use of integral equations for compartmental as well as non-compartmental systems has been discussed by Branson (19-20), Stephenson (54), Zierler (58), Bergner (6), Perl (39) and others.

An interesting treatment of systems using a combination of integral and differential equations has recently been discussed by Hart (32). Under his scheme, if a system is accessible at n sites for the introduction of initial conditions *and* for measurements it is possible, through observation of n^2 response functions to derive the transfer functions (the transform of the weighting functions) or 'first entry' functions between all the sites.

The solution of the linear, constant coefficient differential equations may also be derived through the use of Laplace transforms. The matrix and transform methods are equivalent although each has advantages for special purposes. Treatments of compartment systems using Laplace transforms and other methods have been discussed (12,15,41,42,45,51). Treatments of special compartmental systems have also been published (35,46,47,48,50,52).

Special publications dealing with compartmental analysis are contained in the Annals of the New York Academy of Sciences (1963), the Oak Ridge symposium proceedings on Compartments, Pools and Spaces, edited by Lushbaugh and Bergner (1967) and the books by Rescigno and Segre (42) and Sheppard (51).

The mathematical treatment of non-linear differential equations is limited either to very simple systems or to very special ones. They can be solved, however, numerically on computers. Numerical stepwise solutions using the Euler, Runge-Kutta or other predictor-corrector methods are most commonly used (25). Solutions may also be obtained with analog computers. The convolution integral approach, a powerful tool for linear time invariant system analysis cannot be applied to non-linear systems.

THE INVERSE PROBLEM

So far we discussed the solution of known systems, i.e., the differential equations describing the system and the values of the coefficients are given. The problem one faces in practice is the inverse. Estimates of the solution are available through experimental observations and it is necessary to deduce the structure of the system and to derive estimates for the values of the parameters. It is important to realize that there is *no unique solution* for this problem. Even under ideal conditions there is an infinity of system configurations and sets of parameter values to fit any particular solution. On the other hand, there is also an infinity of systems that will not fit the data. The approach that one follows is to postulate a "simplest" system, consistent with available information, theory, established concepts and preconceived notions, and subject it to rigorous testing. This, in essence, is model building (Berman, 1963a).

A most important ingredient in modeling is the *purpose*. If the purpose is not clear the value of a model is questionable. Some purposes for models are:

1. A formal description of data for conciseness and interpolation.
2. A description of a system in terms of parameters to permit classification based on parameter values.
3. Derive some special measure of interest (for example, pool size, rate of synthesis, clearance rate, etc.).
4. Integrate information and concepts about a system into a common formal framework.
5. Predict responses of systems to various perturbations.
6. Derive the mechanistic principles underlying the behavior of the system.

It is quite clear that for some purposes the nature of the model is immaterial whereas for others it is very critical. Given a purpose it is possible to test the sensitivity of the model with respect to the purpose. When various models can satisfy the same purpose, it becomes a matter of convenience or efficiency as to which is selected.

It is theoretically possible to proceed formally from a set of solutions to parameter values of a known set of linear differential equations with time invariant coefficients. If λ is the matrix of coefficients, A the matrix of its eigenvectors and α the diagonal matrix of its eigenvalues,

$$\lambda = A\alpha A^{-1}$$

In practice, however, it is not possible to take advantage of this relation frequently. The problems that arise may be enumerated as follows:

1. The nature of the system studied is not fully known.
2. The sites available for measurements are restricted.
3. The times when measurements can be made are restricted.
4. The data contain fluctuations.
5. The number of studies that are possible to perform on a single individual is limited.
6. The system is studied by numerous experimental techniques and the data obtained with the various techniques are interrelated in a complicated way.

To deal with these problems, hypotheses and assumptions must be introduced and rigorously tested against the data. The assumptions must also be evaluated with respect to the purpose of the modeling, since assumptions critical with respect to one purpose may not be critical with respect to another.

DATA FITTING

The usual procedure for testing a model is to propose some minimal structure, assign initial values for the parameters, obtain a solution and compare it to the observations. The parameter values are then adjusted until some "best" fit to the data is obtained (13,17,18). The sum of squares of deviation between the calculated and observed values is most commonly used as a criterion for a "best" fit (57). Various convergence procedures are available to get to a "best" fit. Some involve the calculation of partial derivatives of the data with respect to each parameter being adjusted. Others use a steepest descent method, a simplex method, etc. (see (56), for a survey of non-linear optimization techniques). All are iterative methods and none guarantee convergence. In the neighborhood of a least squares fit it is possible to derive a covariance matrix which serves as a measure of distribution for the parameter values, and a measure for goodness of fit such as a chi-squared test (1). From these tests it is possible to judge whether the fit is unique and consistent. Very large variances suggest that the data are inadequate to define all the parameters of the model, in which case further assumptions or constraints must be introduced in the model or new data must be obtained. Inconsistency of a fit results from systematic or unlikely deviations, and suggests that the chosen model is inappropriate or that the data are not what they are believed to be. Unfortunately, these two possibilities cannot be differentiated, and either the model must be modified or the data must be carefully reexamined to resolve the inconsistency. An artificial inconsistency may arise due to a computational failure to reach a "best" fit. This, too, must be studied carefully although there are no fool-proof procedures to ensure convergence to an absolute minimum in the sums of squares surface.

A least squares fit also provides a covariance matrix that contains information about the reliability of the parameter estimates and interrelations in their uncertainties. This information, however, is based on linear multi-variate analysis and extension to non-linear analysis can introduce serious distortions.

In short, for rigorous model testing sophisticated statistics and computational techniques must be applied—with no assurance of success.

STRATEGY FOR EXPERIMENTATION

As pointed out earlier, a tracer experiment performed on a steady state nonlinear system reflects only part of the system. It has been shown by Berman and Schoenfeld (12) that for *n* compartments there is a minimal amount of data that must be available to determine all the parameter values in the model. Such data may be collected as observations in the *n* compartments due to a single set of initial conditions or as observations in one compartment due to multiple sets of initial conditions in all compartments (50). Data, however, may also come from other sources, such as steady state tracer measurements, a-priori information about the existence or non-existence of certain pathways, specific activity values for some compartment, etc. Most frequently the data are a combination of these. To test whether a model can be uniquely determined from all the fragmentary data available it is necessary to simulate all the data with a model and adjust the model parameters until an overall least squares fit is obtained. From the convariance matrix associated with the fit it is then possible to determine whether the data are adequate.

Theoretically, tracer experiments performed beyond the minimum required to define a model should only improve the precision of the estimate for the parameter values. This, however, assumes that the model is correctly known. Since in practice this is not always the case, additional experiments contain new information when they disclose inconsistencies. Once a model is completely tested it is a composite of all the information that can be extracted about the system with the experimental techniques employed.

Tracer experiments on a steady state system contain limited information. To extract more information it is necessary either to perturb the system to new steady states and repeat the tracer experiments at the new states, or to study it during non-steady states. The latter is a difficult approach since one usually deals with relatively complex unknown structures and the likelihood of coming up with a good model is small. It also imposes heavy demands on the quantity and quality of the data. Tracer studies at various steady states have the advantage that the model is simpler and remains the same for various perturbations. Changes in parameter values for various steady states suggest functional relations between the tracer parameters (λ_{ij}) and tracer concentrations. This, in turn, provides an insight into the nature of various interactions and leads to systematic refinements of the model.

Tracer studies at multiple steady states is especially rich in information if the sites of perturbation in the system are known and few. It was shown by Berman (7) how limited tracer data from each of a number of perturbation experiments can jointly lead to unique solutions.

The study of a homogeneous population of individuals may be viewed as a special case of multiple perturbations. If there is such a thing as a population mean, each individual represents a deviation from that mean. The covariance matrix of the deviations shows which parameters are correlated and which are independent—information that cannot be extracted from a tracer experiment on one individual.

Ultimately one wishes to know how a system or sub-system relates to others. The biological organism is full of coupled systems that regulate various functions in the body. For example, glucose is controlled by insulin, insulin by glucose, lipids by insulin, etc. The simultaneous study of glucose and insulin in an individual contains more information than separate studies of glucose and insulin done on the same individual at different times, and even more information than a glucose study in one individual and insulin study in another. Similarly, the study of multiple subsystems in one individual contains information not readily obtainable from populations of individuals each studied for only one of the subsystems. To perform multiple experiments on a single individual, however, is difficult especially if the individual is normal. Such experiments tend to become extensive, require the participation of investigators in various disciplines and may result in a relatively large radiation burden on the individual studied if radioactive isotopes are used. Well organized and integrated teamwork in the execution of the experiment and in the data analysis and modeling is also required.

DATA PROCESSING

To deal with complex systems it is essential to have a large, modern computer. It is further essential that the computer be able to simulate all experimental conditions, observations, constraints and changes jointly so that all independent and correlated information may be brought to bear simultaneously in testing a model. Furthermore, it is necessary that proper statistical weights be assigned to the observations so that each will contribute to the total information content in accordance with its precision. The derived model should be physically realizable and parameter values should stay within the resolution limits of the data. Constraints and uncertainties arising from the experimental techniques should also be incorporated in the simulation.

There are many computer programs for data fitting. There are very few, however, designed with a global view towards kinetic analysis and modeling. The SAAM (Simulation, Analysis and Modeling) computer program that we have been developing over a number of years (9,13,14) has some of the desired features and permits the integration of various sources of information about a system into a common simulation and data fitting process. SAAM was developed with the intent to enable a mathematically unsophisticated physiologist or a biophysicist without extensive programming experience and knowledge of numerical analysis or statistics to converse with the program and do sophisticated modeling. To achieve this, a formalism for model building and a language had to be developed. But much still remains to be done in this area.

The integration of many sources of information in the study of normal and abnormal physiological states for several subsystems leads to large, complex models. A number of problems arise from such models.

Convergence: Most models are non-linear in relation to the data. This means that iterative procedures must be employed to converge to a best fit. The convergence surface is highly irregular and it is possible to get trapped in numerous local minima, unless 'good' initial estimates are available. It is also possible that multiple minima exist which fit the data equally well. If a "good" fit cannot be derived it is impossible to tell, without extensive testing, whether the model is inconsistent or the solution is trapped in some local minimum. Unfortunately, when dealing with large systems intuition fails and one is forced to resort to brute force methods for testing. This raises computing costs considerably.

Computing time: Large, non-linear sets of differential equations require stepwise solutions. When the model contains parameters that reflect both slow and rapid processes, the step size is controlled by the rapid processes whereas the number of steps is controlled by the slow processes (duration of the experiment). Coupled with the facts that at each step the number of computations go up with model size and the rate of convergence goes down, the computing time and costs can become prohibitive even on the largest and most efficient computers available. Analog computers, hybrid systems or parallel processes (like ILLIVAC IV), with man-machine interaction may be required to deal with these problems.

Uniqueness: When a large model evolves from many sources of information, it is too large to determine uniquely parameter values for a single individual on whom limited data are available. One way to deal with this is to fit the data subject to the constraint that the individual belong to a population for which parameter mean values and a covariance matrix are known. The values derived for the individual will then be 'as close' to the mean of the population as his data will permit. To derive population values, however, a large number of individuals must be studied extensively. To compile population values for various systems in various physiological states therefore is an undertaking that will require the participation of many investigators over a considerable span of time.

Because of the many assumptions and limitations of modeling and data analysis, kinetic modeling can play only a limited role. It is most powerful in testing hypotheses for consistency. It cannot, however, serve as proof for the reality of a model when it is consistent. The intuition and prejudice of the investigator remain the rationale in the selection of a model and in proclaiming its 'truth'.

Example

There are many examples of the use of compartmental analysis in metabolic studies in calcium, glucose, fatty acids, insulin, etc. Some of these are listed by Berman and Weiss (14) and by Baker (3). Let me use iodine kinetics as an example. We have been modeling this system for a number of years during which time the model underwent many transformations. Most model changes occurred early and relatively few in recent years.

A typical iodine experiment is performed by injecting into the plasma labeled iodine in the form of iodide, thyroxine (T_4) or triiodothyronine (T_3), and following its course in plasma, thyroid, urine and feces. Tracee levels of PBI, daily urine, etc., are also measured. Data from studies in normal man and in various abnormal thyroid states are shown in figure 2.

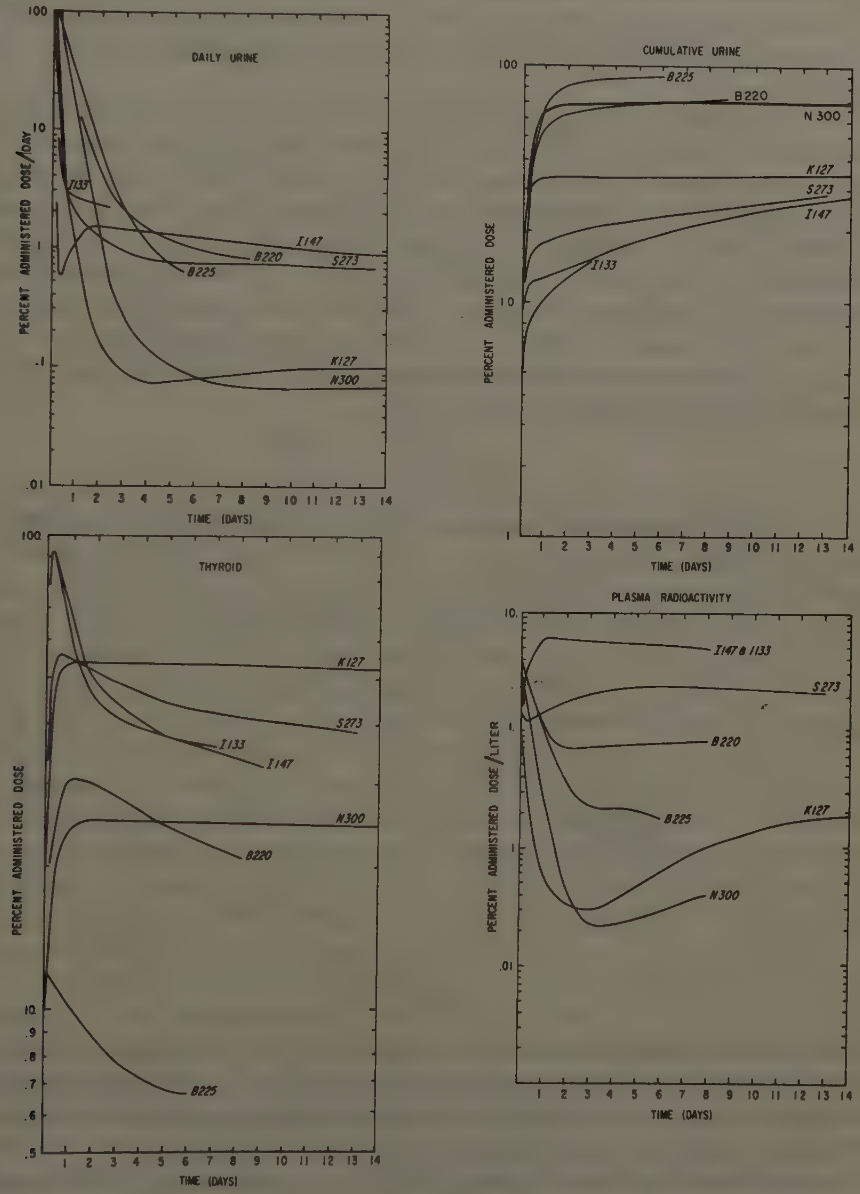


Fig. 2

In the first attempt to fit the data several compartmental models suggested in the literature were tested (16,22,43,44). A total of about 150 kinetic studies in various normal and abnormal states of thyroid function were tested and it was soon found that many were inconsistent with the model. Although certain 3 or 4 compartment models could be made to fit some of the studies, the requirement that a *single, universal model fit them all* forced us to reject them.

Various models were then tested under the constraint that they relate as much as possible to known physiology, with structure no finer than the resolution of the data justify. Limitations in resolution means that states that interact faster than the data can resolve get lumped with slower interacting states into single compartments.

After some extensive testing, a model, shown in figure 3, that fits practically all our studies was evolved (11). The model contains four subsystems: plasma iodide, thyroid, extra-thyroidal T_4 and extra-thyroidal T_3. The thyroid subsystem required several new physiological features: 1) that it be separated into 3 phases; an initial rapid release phase of hormone (~1 day), a delay phase (1-4 days) during which no hormone is released, and a storage phase with a slow rate of hormone release (~100 days), 2) significant amount of T_3 are released by the thyroid and 3) T_3 and T_4 may be released in different ratios in the early phases compared to the late ones. The model also required iodide release from the thyroid. Some of these features still require physiological confirmation.

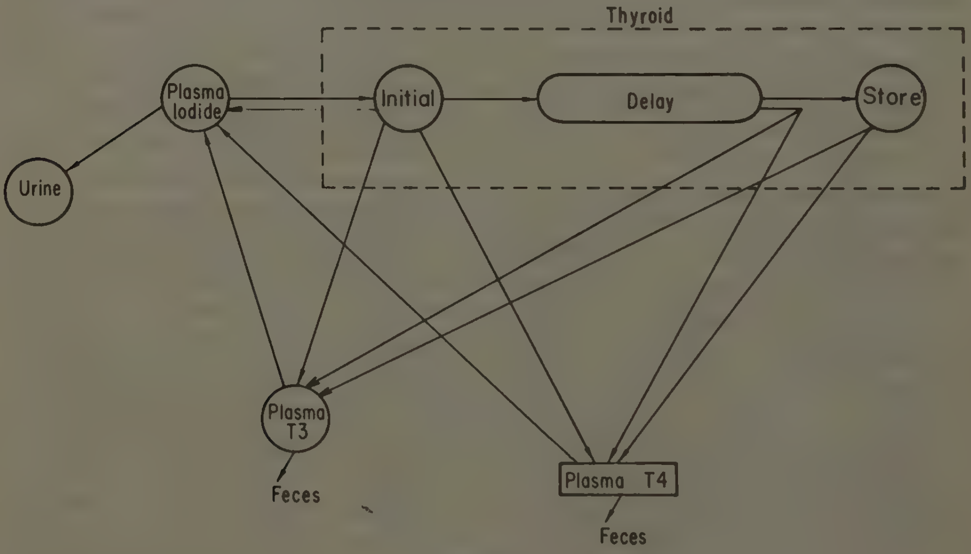


Fig. 3

Finer structure has recently been proposed for the T_4 subsystem (24,36,38, 47). According to these investigators the plasma, liver and extravascular (and tissue?) spaces can be resolved. Incorporation of this subsystem and a similar one for T_3 based on our own work leads to the model shown in figure 4.

Compartment 1 represents plasma iodide, compartments 2, 5, 11 are states within the thyroid. Compartments 9, 4, and 14 are extra-thyroidal thyroxine and

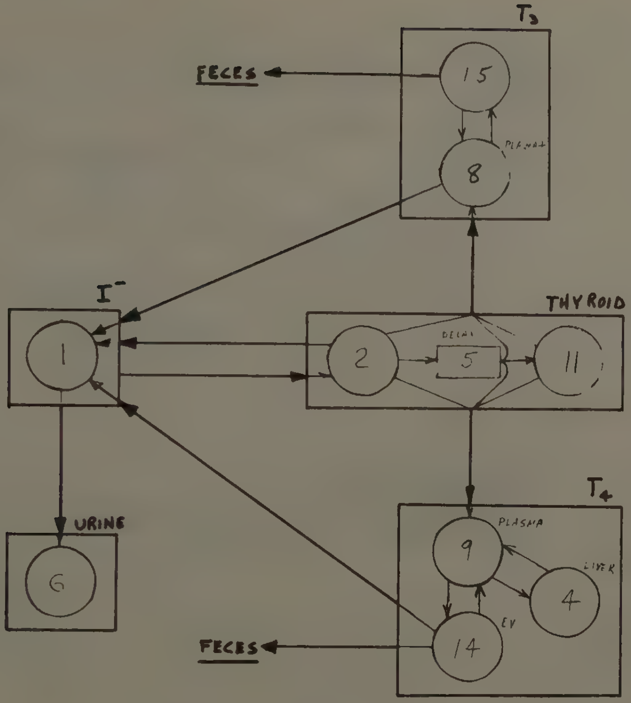


Fig. 4

compartments 8 and 15 are extra-thyroidal triiodothyronine (T_3). Compartment 6 is urine. Compartment 4 of the T_4 subsystem is identified as liver.

Figure 5 shows the superposition of the experimental constraints on the model in simulating the measurements. The triangles indicate sites of measurement and the dotted lines represent fractions of the various compartments contributing to each measurement. For example, component 3 represents measurements, q_3, over the thyroid and is expressed as $q_3 = \sigma_{31}f_1 + \sigma_{32}f_2 + \sigma_{35}f_5 + \sigma_{3,11}f_{11} + \sigma_{38}f_8 + \sigma_{3,15}f_{15} + \sigma_{39}f_9 + \sigma_{3,14}f_{14}$. σ_{31} represents the fraction of the iodide compartment that is seen by the radiation detector when measuring the thyroid, σ_{38} and σ_{39} are similar fractions for plasma T_3 and T_4, respectively, etc. σ_{32}, σ_{35}, and $\sigma_{3,11}$ represent the fraction of the total thyroid content as seen by the detector and deviations from unity represent standardization corrections. In fitting the data the computer calculates the values of the σ_{ij}. In the case of the thyroid the fitting is subject to the constraint that $\sigma_{32} = \sigma_{35} = \sigma_{3,11}$, since various thyroid components are measured with equal efficiency.

The reciprocal of the fractions σ_{71}, σ_{78} and σ_{79} represent plasma equivalent spaces (23) for compartments 1, 8 and 9, respectively.

Some sub-systems are still not fully resolved. The iodide distribution into various organs is too rapid and inaccessible to permit quantification (Sharney et al 1965b). The known fine structure of hormone synthesis in the thyroid is also excluded from the model although a fair amount of information is available about this (30,40). DeCostre et al (27) have made an effort recently to model this in greater detail. Also excluded from the model are all non-linear aspects of iodine

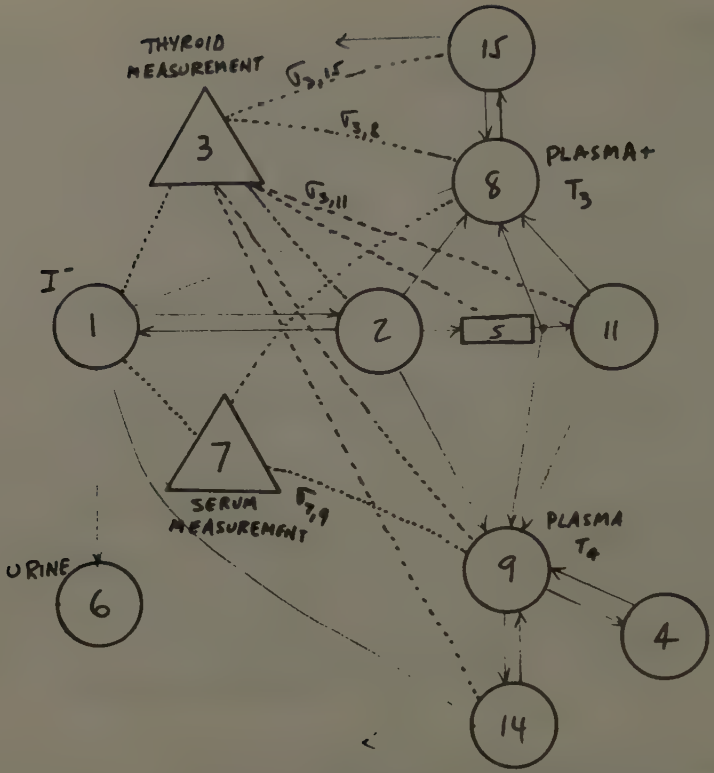


Fig. 5

kinetics (iodide trapping by the thyroid and hormone degradation), the effects produced by abnormal states and radiation treatment, the regulatory aspects of iodine metabolism due to TSH, LATS and other factors (21,26,28,29,37). Also excluded from the model are the binding properties of various iodine compounds to proteins which affect their kinetics and distribution.

It is obvious, that the iodine kinetics system is very complex and even in the steady state the subsystem requires considerable data to determine parameter values. To collect enough data on a single individual is a large undertaking and can be justified only in the context of a research project.

Several techniques may be employed to aid in the resolution of large system. First, subsystems may be studied on populations of individuals and population means derived. These may then be imbedded into the larger models. Secondly, perturbation studies (8) may be performed which jointly can improve on the resolution capabilities of each.

Figure 6 shows a fit of 2 sets of iodine data collected on a patient before and after radiation treatment. Each study by itself is inadequate to define the parameter values. If it is postulated, however, that radiation damage is restricted to a limited number of sites, a unique solution for pre- and post-therapy can be obtained (11).

To do extensive studies on an individual for routine diagnosis is impractical. Large models may, however, be employed to aid in diagnosis if limited data are

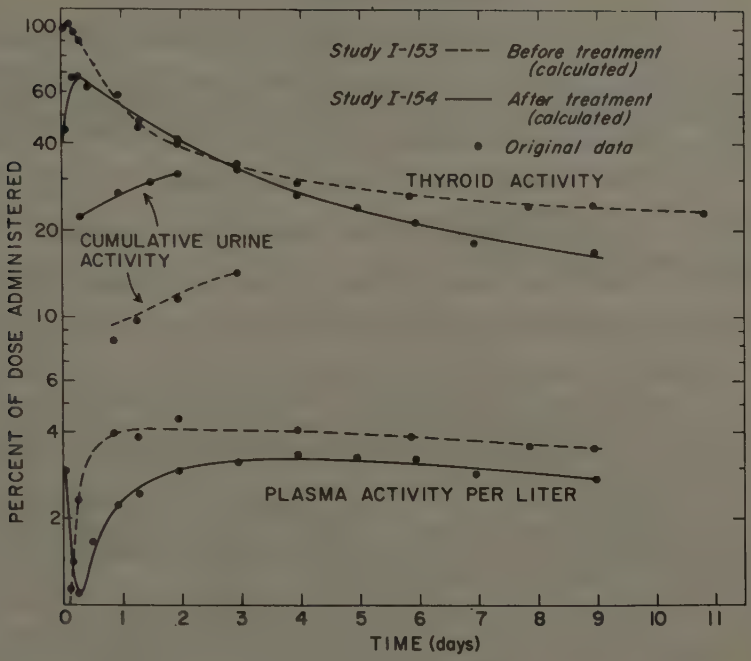


Fig. 6

available. The data are fitted to the model subject to the constraint that an individual be a member of a specified population—unless his data make it unlikely. This involves the fitting of limited data (even one or few observations) to the mean of the population values with an *a priori* distribution governed by the population covariance matrix. Although such a fit does not prove an abnormality it provides a measure of likelihood for encountering the observed values in some specified population.

Summary.

Careful kinetic studies in biological systems eventually lead to extensive models. To quantify these, information may be extracted from many sources and all must be integrated into a single model. Although the treatment of the data using analytical methods can yield useful information in special cases, the dealing with large models, diversity of information and non-linearities, and the necessity to test the model rigorously against the data requires the use of high speed computers and sophisticated techniques in numerical analysis and statistics. It is also important for the modeler to know the biology (physiology, biochemistry, etc.) of the system he studies. All this leads to the requirement of team efforts.

Although large models are useful in understanding complex systems — they are likely to be derived through a combination of studies on populations and individuals. Once population values are defined large models serve as a framework within which the interpretation of limited data in individuals can be evaluated.

REFERENCES

1. Anderson, T. W., 1958. An introduction to multi-variate statistical analysis. John Wiley and Sons, Inc., New York.
2. Annals of the New York Academy of Sciences. 1963. Multi-compartment analysis of tracer experiments. *Annals of the New York Academy of Sciences* **108:** Art. 1, 1-338.
3. Baker, Nome, 1969. The use of computers to study rates of lipid metabolism. *J. Lipid Res.* **10:** 1-24.
4. Bellman, Richard, 1953. Stability theory of differential equations. Chapter 1, McGraw Hill Book Co., Inc., New York.
5. Bergner, P.-E. E., 1965. Exchangeable mass determination without assumption of isotopic equilibrium. *Science* **150:** 1048-1050.
6. Bergner, P.-E. E., 1966. Tracer theory a review. *Isotopes and Radiation Technology* **3:** 245-262.
7. Berman, M., 1963^a. The formulation and testing of models. *Ann. N.Y. Acad. Sci.* **108:** 182-194.
8. Berman, M., 1963^b. A postulate to aid in model building. *J. Theoret. Biol.* **4:** 229-236.
9. Berman, M., 1965. Compartmental analysis in kinetics. In Computers in Biomedical Research, Vol. 2, Chapter 7, Stacy and Waxman, editors, Academic Press, New York.
10. Berman, M., 1968. The application of multicompartmental analysis to problems of clinical medicine. *Ann. Intern. Med.* **68:** 423-448.
11. Berman, M., Hoff, E., Barandes, M., Becker, D. V., Sonenberg, M., Benua, R. and Koutras, D. A., 1968. Iodine kinetics in man—a model. *J. Clin. Endoc. and Metab.* **28:** 1-14.
12. Berman, M. and Schoenfeld, R., 1956. Invariants in experimental data on linear kinetics and the formulation of models. *J. Appl. Phys.* **27:** 1361-1370.
13. Berman, M., Shahn, E., and Weiss, M. F., 1962. The routine fitting of kinetic data to models—a mathematical formalism for digital computers. *Biophys. J.* **2:** 275.
14. Berman, M. and Weiss, M. F., 1967. Saam Manual, NIH PHS Publication, 1703.
15. Berman, M., Weiss, M. F., and Shahn, E., 1962. Some formal approaches to the analysis of kinetic data in terms of linear compartmental systems *Biophys. J.* **2:** 289.
16. Berson, S. A. and Yalow, R. S., 1954. Quantitative aspects of iodine metabolism. The exchangeable organic iodine pool, and the rates of thyroidal secretion, peripheral degradation and fecal excretion of endogenously synthesized organically bound iodine. *J. Clin. Invest.* **33:** 1533-1552.
17. Box, G. E. P., 1960. Fitting empirical data. *Annals of the N.Y. Acad. of Sciences* **86:** 792-816.
18. Box, G. E. P. and Hunter, W. G., 1965. The experimental study of physical mechanisms. *Technometrics* **7:** 23-42.
19. Branson, H. 1946. A mathematical description of metabolizing systems I. *Bull. Math. Biophy.* **8:** 159-165.
20. Branson, H., 1947. A mathematical description of metabolizing systems II. *Bull. Math. Biophy.* **9:** 93-98.
21. Brown-Grant, K., 1967. The control of thyroid secretion. *J. Clin. Path., Suppl.* **20:** 327-332.
22. Brownell, G. L., 1951. Analysis of techniques for the determination of thyroid function with radioiodine. *J. Clin. Endocrin.* **9:** 1095-1105.
23. Brownell, G. L., Berman, M., and Robertson, J. S., 1968. Nomenclature for tracer kinetics. *Intenat. J. Appl. Radiat. and Isotopes* **19:** 249-262.
24. Cavalieri, Ralph R. and Searle, Gilbert, L., 1966. The kinetics of distribution between plasma and liver of I-131 labeled L- thyroxine in man. Observations of subjects with normal and decreased serum thyroxine-binding globulin. *J. Clin. Invest.* **45:** 939-949.
25. Collatz, L., 1960. The numerical treatment of differential equations. Springer-Verlag.
26. Danziger, L. and Elmergreen, G. L., 1956. The thyroid-pituitary homeostatic mechanism. *Bull. Math. Biophys.* **18:** 1-13.
27. DeCostre, P. L., Phair, R. D., Dingwell, I. W. and Degroot, L. J., 1968. A thyroid model and its analysis of computer. Abstract, Annual Meeting American Thyroid Assoc., Washington, D.C.
28. Distefano, J. J. III., Hypothalamic and rate feedback in the thyroid hormone regulation system. An hypothesis. *Bull. Math. Biophysics* **31:** 233-46.

29. Distefano, Joseph J. III and Stear, E. B., 1968. Neuro-endocrine control of thyroid secretion in living systems. A feedback control system model. *Bull. Math. Biophysics* **30:** 3-26.
30. Ermans, A. M. and Camus, M., 1966. Research concerning the influence of acute exposure to cold on the thyroid function. Arctic Aeromedical Lab. Report AAL-TR-66-7.
31. Gurpide, Erlio, Mann, J. and Liberman, S., 1963. Analysis of open systems of multiple pools by administration of tracers at constant rate as a single dose as illustrated by problems involving steroid hormones. *J. Clin. Endoc. and Metab.* **23, 11:** 1155-1176.
32. Hart, H. E., 1967. Analysis of tracer experiments. VIII integro-differential equation treatment of partly accessible, partly injectable multicompartment systems. *Bull. Math. Biophy.* **29:** 319-333.
33. Hearon, J. Z., 1963. Theorems on linear systems. *N.Y. Academy of Sciences* **108:** 36-38.
34. Lusbaugh. C. C., Bergner, P. E. E., Editors, Compartments, pools and spaces. USAEC Symposium Series 11, Oak Ridge.
35. Matthews, C. M. E., 1957. The theory of tracer experiments with 131-I labelled plasma proteins. *Phys. Med. Biol.* **2:** 36-53.
36. Nicoloff, J. T. and Dowling, Thomas J., 1968. Estimation of thyroxine distribution in man. *J. Clin. Invest.* **47:** 26-37.
37. Norwich. K. H. and Reiter, R., 1965. Homeostatic control of thyroxin concentration expressed by a set of linear differential equations. *Bull. Math. Biophys.* **27:** 133-144.
38. Oppenheimer, Jack H., Bernstein, Gerald, and Hasen, Julian, 1967. Estimation of rapidly exchangeable cellular thyroxine from the plasma disappearance curves of simultaneously administered thyroxine-1131 and albumin-1125. *J. Clin. Invest.* **46:** 762-77.
39. Perl, W., Effros, R. M., and Chinard, F. P., In press: The indicator equivalence theorem for input rates and regional masses in multi-inlet steady state systems with partially labeled input. *J. Theoret. Biol.*
40. Rall, J. E., Robbins, J., and Lewallen, C. G., 1964. The thyroid. In The Hormones, Vol. 5, Chapter 3, Pincus, Ed., Academic Press, N.Y.
41. Rescigno, Aldo, 1954. A contribution to the theory of tracer methods. *Biochimica et Biophysica Acta* **15:** 340-344.
42. Rescigno, A. and Segre, G., 1966. Drug and tracer kinetics. Blaisdell Pub. Co., Waltham, Mass.
43. Riggs, D. S., 1952. Quantitative aspects of iodine metabolism in man. *Pharmacol. Revs.* **4:** 284.
44. Riviere, R., Comar, D., and Kellersohn, C., 1965 A comparative analysis of the curves of the iodine specific activity determined in man throughout long-period studies, in the different compartments of the thyroid metabolic system. In Current topics in thyroid research. Academic Press, New York.
45. Robertson, James S., 1962. Mathematical treatment of uptake and release of indicator substances in relation to flow analysis in tissues and organs. In Am. Physiol. Society Handbook of Physiol., Vol. II, Chapt. D2, Sec. 2.
46. Segre, Giorgio., 1962. Applications of the transfer functions to pharmacology. *Ann. N.Y. Acad. of Sci.* **96:** 913-938.
47. Sharney, Lena, Segal, R. L., Dumont, A. E., Girolami, A. and Silver, S., 1965. Studies in iodine metabolism III. Three pool systems of extrathyroidal thyroxine kinetics. *J. of the Mount Sinai Hospital* **32:** 396-420.
48. Sharney, L. R., Segal, L., Witte, M. H., Girolami, A. and Beck, R., 1965. Studies in iodide metabolism II. Three-pool systems of iodode kinetics. *J. of the Mount Siani Hospital* **32:** 375-395.
49. Sharney, L., Wasserman, L. R., Gevirtz, N. R., Schwartz, L., and Tendler, D., 1965. Multiple-pool analysis in tracer studies of metabolic kinetics. *J. Mt. Sinai Hosp.* **32:** 201-261.
50. Sheppard, C. W. and Householder, A. S., 1951. The mathematical basis of the interpretation of tracer experiments in closed steady state systems. *J. Appl. Phys.* **22:** 510-520.
51. Sheppard, C. W., 1961. Principles of the tracer method. John Wiley and Sons, Inc.
52. Skinner, S. M., Clark, R. E., Baker, N. and Shipley, R. A., 1959. Complete solution of the three compartment model in steady state after single injection of radio-active tracer. *Am. J. Physiol.* **196:** 238-244.
53. Solomon, A. K., 1960. Compartmental methods of kinetic analysis. Mineral metab., Vol. 1, Part A, Academic Press.

54. Stephenson, J. L., 1960. Theory of transport in linear biological systems I. Fundamental integral equation. *Bull. Math. Biophysics* **22:** 1-17.
55. Stephenson, John L., 1960. Theory of transport in linear biological systems II. Multiflux problems. *Bull. Math. Biophys.* **22:** 113-138.
56. Swann, W. H., 1969. A survey of non-linear optimization techniques. FEBS letters, Vol. 2, Supplement, 39-55.
57. Whittaker, E. and Robinson, G., 1944. The calculus of observations. Blackie and Son, Limited. London, England.
58. Zierler, Kenneth L., 1963. Effects of circulatory beds on tracer experiments or non-compartmental analysis. *Ann. N.Y. Acad. Sciences* **108:** 106-116.

RESOLUTION RECOVERY IN RADIOISOTOPE SCANS

Stephen M. Pizer

Scan* processing using a digital computer promises to increase the diagnostic information in radioisotope scans. It can accomplish this both by providing scans in a form well matched to man's perceptual abilities and by introducing added information from knowledge of the imaging spread function and the statistical distribution of "noise" and from *a priori* knowledge about the object that was scanned.

The object of scan processing methods can be divided into three categories smoothing, resolution recovery, and transformation for display. Smoothing, which reduces statistical fluctuations (noise) and transformation for display, e.g. contour display, contrast enhancement, volumetric display, have as their purpose providing scans in the optimum form mentioned above. Smoothing can also be thought of as part of the attempt to recover the original activity distribution, which is known to be smoothly varying (except perhaps for discrete steps), unlike the scan itself which includes statistical fluctuations. The necessity for such smoothing decreases as the number of events in a scan increases, lowering the relative level of the statistical fluctuations.

Another type of processing which is necessary if one is trying to estimate the original activity distribution is recovery from the blurring caused by imaging and previous processing. The problem with such resolution recovery processes is that they enhance statistical fluctuations as they enhance true variations in intensity. However, as the statistical fluctuations are lower this resolution recovery can be more and more accurately accomplished.

Since new scanning devices are becoming available which provide increased counts per unit time, smoothing is becoming less necessary on the one hand, and resolution recovery is becoming more effective on the other. Thus, this paper will focus on the problem of resolution recovery in radioisotope scans. Such a focus should not be taken as implying that the problem of transformation for display is unimportant or solved. On the contrary, the success of any processing strongly depends on the display methods used to present the results of the processing.

---

* Throughout this paper "scanning" will refer to any process which records a radioisotope distribution. "Scanner" will refer to the machine which records this distribution and "scan" will refer to the output of the scanner.

The characteristic which distinguishes radioisotope scans from most other pictures is the fact that the "noise level", i.e. the level of statistical fluctuations, at any point in the scan is a function of the intensity at that point. This results from the fact that scans are quantum limited, i.e. the number of events in a picture is small. Thus the Poisson distributed statistical fluctuations are a significant source of noise.

The fact that the noise varies over the scan implies that any processing routine which deals with this noise should vary the processing as a function of the local intensity distribution about the point at which the result is being computed. With respect to resolution recovery, the argument can be further supported as follows. In areas of low event density, the statistical resolution, i.e. the best resolution obtainable for that event density and contrast with perfect imaging, is poor, and thus geometric information is low, because the relative statistical fluctuations are high. Thus any attempt to recover resolution must be very limited, because the scan in that area includes little geometric information which can be recovered or to look at it another way, any attempt to recover much resolution will enhance the large statistical fluctuations. In areas of high event density, however, there is much geometric information available because of low relative statistical fluctuations. Thus an attempt to recover resolution can be quite successful.

One should note that the argument above states that the processing should depend on the very thing we are trying to find, the intensity distribution. This apparent circularity can be resolved by noting that the event density distribution which is input to the process is itself an estimate of the intensity distribution and its values can be used, albeit with care, to set the variation of the processing.

Resolution recovery methods can be divided into three types: Deconvolution by filtering, dot or density shifting, and iterative unfolding. These as well as specific processing routines to implement them will be discussed in the following.

Deconvolution by Filtering

The object of such a process is to find a spread function inverse to the imaging spread function (Fig. 1) and to apply some approximation of this inverse spread function or filter to the image. The approximation can be computed in number of different ways, including methods using Fourier analysis and methods involving spatial derivatives of the intensity distribution. Furthermore, the process can be applied either by convolution in the space domain or by transformation into the frequency domain by fast Fourier transform, multiplication, and transformation back to the space domain. Such methods all have the same shortcoming—there must be a compromise between a filter which recovers much resolution, useful in the more intense areas of the scan, and one which does not overly enhance statistical fluctuations, necessary in the less intense areas of the scan.

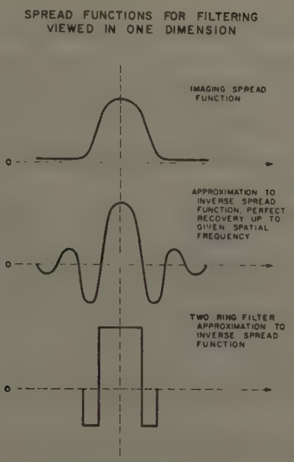

SPREAD FUNCTIONS FOR FILTERING
VIEWED IN ONE DIMENSION
IMAGING SPREAD FUNCTION
APPROXIMATION TO INVERSE SPREAD FUNCTION. PERFECT RECOVERY UP TO GIVEN SPATIAL FREQUENCY
TWO RING FILTER APPROXIMATION TO INVERSE SPREAD FUNCTION
0
0
0

Fig. 1. Imaging and Inverse Spread Functions   Fig. 2. Fixed Filtering with Two Ring Filter

To solve this problem one should design a process where the filter varies as a function of the local intensity distribution. A useful criterion for this variation is that the signal to noise ratio of the output should be constant over the scan.

The necessity of computing a different filter for each point at which the output is to be computed can make the processing a very time consuming operation. To avoid this problem, one must choose a form for the filter with few parameters. Such a form is the two-ring filter (Fig. 1, Fig. 2.) where the inverse spread function is approximated by two uniform areas, a circle and a concentric ring with the same area, where the filter value is a positive constant (α) in the inner circle and a negative constant ($-\beta$) in the outer ring. The values of these constants can be computed as a function of the radius of the circle. The radius is chosen at each computation point to produce a fixed signal to noise ratio by satisfying a relation between the number of events in the inside circle and the number in the outer ring:

$$Q \leq \frac{[A(N_1 - N_2) + N_1]^2}{(A+1)^2 N_1 + A^2 N_2},$$

where $A = \dfrac{2\sigma^2}{R_1{}^2} + \dfrac{1}{2}$

and N_1 and N_2 are the number of dots in the inner circle and outer ring respectively,

R_1 is the radius of the inner circle,

Q is the signal to noise ratio desired,

σ^2 is the variance of the imaging spread function (assumed to be gaussian)

In initial work with variable filtering, the variation of the filter has only depended on the number of dots in the inner circle, while assuming for the choice of filter that the number of dots in the outer circle is the same(3). Even using this simplification, results are promising (Fig. 5, C vs B). Even better results are anticipated when the variation depends on the numbers of dots in both areas.

Dot or Density Shifting

Dot (event) shifting conceives of the blurring due to imaging as a shifting of the dots (detected events) from the points at which they were produced to those at which they were detected. It attempts the inverse operation by shifting the dots to their most likely origins. Each dot should be shifted by an amount and in a direction determined by the local dot density distribution. If we take the example of a high contrast boundary, we can easily see that for a dot recorded in the low density field but close to the boundary the probability it was caused by an event from the high density field is increased in accordance with the area of the higher density field seen by the detector at that time. The best estimate for this event's origin should therefore be shifted from the dot toward the boundary (Fig. 3). More correctly, the process of dot shifting should spread the dots about their shifted positions according to a distribution reflecting the uncertainty of that shift. Since the uncertainty depends on the local dot distribution, the spread should vary over the scan.

Dot shifting with spreading can be extended to be applied to intensity (i.e. dot density) values. In this form it will be referred to as density shifting.

The question of what is the most likely origin of an event will now be discussed. If the events are considered independently, we can take the most likely origin to be the centroid (expected value) of the intensity distribution obtained by weighting an estimate of the local density distribution by the imaging spread function centered at the dot or density in question:

$$(x', y') = (x, y) - \frac{\int\int(u, v)F(u, v)C(x - u, y - v)du\, dv}{\int\int F(u, v)C(x - u, y - v)du\, dv},$$

where (x', y') and (x, y) are the new and old positions respectively of the dot being shifted, F is the imaging spread function, and C is an estimate of the dot density distribution.

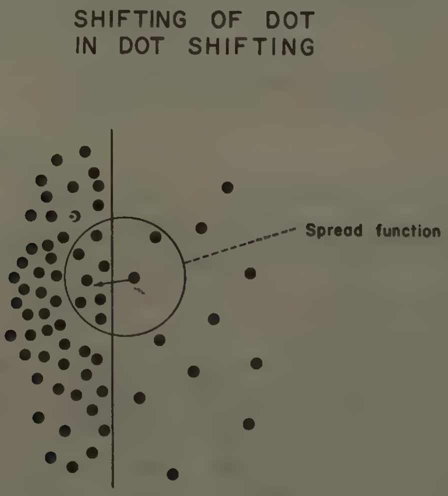


Fig. 3. Shifting of Dot in Dot Shifting

The fact is, of course, that the dots are not independent—they reflect the activity distribution which was scanned. Thus the choice of the most likely origins is more correctly obtained from some multivariate distribution. However, such a procedure quickly becomes computationally unattractive, and so far we have chosen to accept the approximation given by treating the dots or densities independently.

We have so far accomplished the spreading by doing a variable spatial averaging (5) on the shifted unspread dots. In variable spatial averaging, the averaging area about any computational point is varied so as to obtain a fixed signal to noise ratio of the output. This is accomplished by finding the distance, R, from the point to the N^{th} closest event, where N is fixed, and using as the dot density estimate $(N\text{-}1)/\pi R^2$. For typical results using dot shifting followed by variable averaging, see Fig. 5, D vs B.

Dot or density shifting has the advantage over all the filtering methods we have tried that it enhances statistical fluctuations less for the amount of resolution recovery it does. It has the disadvantage that it will not completely recover the resolution of a scan, as filtering can in theory if the scan has no statistical fluctuations. Furthermore with dot or density shifting one cannot choose the signal to noise ratio flexibly as with variable filtering.

Iterative Unfolding (6)

Iterative unfolding involves approximating iteratively the distribution which when convolved with the imaging spread function produces the image being processed. It accomplishes this by modifying the present estimate at a point by adding an approximation of the difference between the correct answer and the present estimate of that point. It approximates this difference by computing the difference between the value at that point of the blurred version of the correct answer and the blurred version of the present estimate, where the blurred version of the correct answer is the image distribution itself and the blurred version of the present estimate is computed by convolving it with the imaging spread function (see Fig. 4.). One can gain convergence speed at the expense of having the process converge for fewer spread functions by multiplying the modifying difference by a constant greater than 1. Similarly one can make the process converge for more spread functions but more slowly by using a positive constant less than 1. The following equation describes the process:

$$C^{(m)}(x, y) = C^{(m-1)}(x, y) + \beta(C^{(0)}(x, y) - \iint F(u, v)C^{(m-1)}(x - u, y - v)du\, dv),$$

where $C^{(o)}$ is the image distribution, $C^{(m)}$ is the result of the m^{th} iteration, F is the imaging spread function, and β is the multiplicative constant.

The iteration is stopped when the difference between the image and the result of convolving the imaging spread function with the present estimate is small enough according to some statistically defined criterion. Thus this method has the advantage at least in theory that it allows one to control the closeness of the approximation desired and thus the enhancement of statistical fluctuations. However, it can have convergence problems and if convergent the *result* may be too

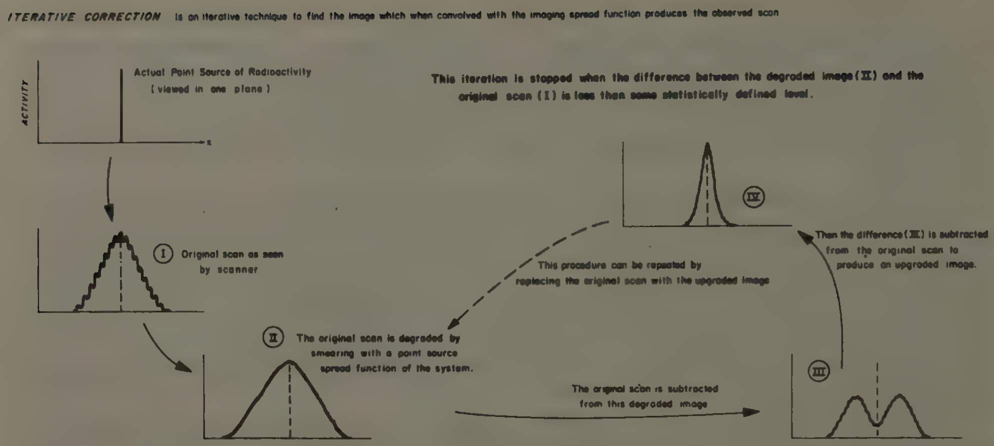
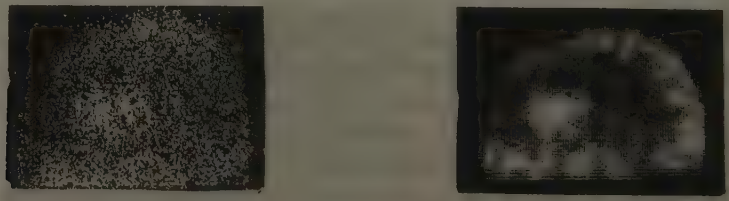

ITERATIVE CORRECTION is an iterative technique to find the image which when convolved with the imaging spread function produces the observed scan
ACTIVITY
x
Actual Point Source of Radioactivity (viewed in one plane)
This iteration is stopped when the difference between the degraded image (II) and the original scan (I) is less than some statistically defined level.
IV
I Original scan as seen by scanner
This procedure can be repeated by replacing the original scan with the upgraded image
Then the difference (III) is subtracted from the original scan to produce an upgraded image.
II The original scan is degraded by smearing with a point source spread function of the system.
The original scan is subtracted from this degraded image
III

Fig. 4. Iterative Unfolding

sensitive to fluctuations. Most probably, its problems are due to the fact that the process does not vary with the local characteristics of the image distribution. How to modify it to do so is not obvious. In any case, it is a very time consuming process and thus is unlikely to be preferred to variable filtering or density shifting.

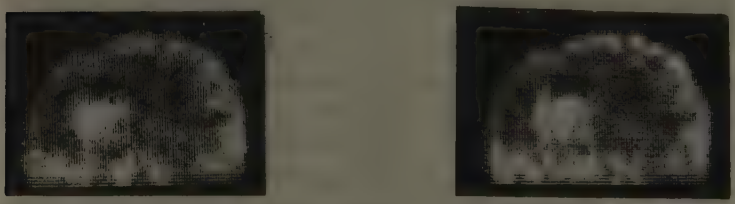
Improvement of resolution of scanners

It may be useful to discuss here the relation between resolution improvement by processing and resolution improvement by designing scanners which trade off some of the gain in sensitivity being achieved now and likely to continue in the future for improvement in resolution. The question of the tradeoff has been treated for the case in which no resolution recovery is to be attempted (7). In the case



A. Dot Picture

B. Variable Average



C. Variable Filter

D. Variable Average of dot shifted picture

Fig. 5. Processing of Human Brain Scan

where resolution recovery will be attempted, it still should be noted that to achieve the tradeoff recommended in (7), resolution improvement of the scanner for that purpose is to be preferred to resolution recovery by processing. However processing to improve resolution does have an advantage over instrument improvement that the processing can vary over the scan whereas present scanners, at least, cannot treat areas differently as a function of the local characteristics of the distribution being scanned.

The above arguments lead one to believe that the optimum tradeoff between sensitivity and resolution for scanners may be shifted slightly toward higher sensitivity if processing for resolution recovery is to be used. Precise calculations of the extent of this shift have not yet been done.

Conclusion

Processing to achieve resolution recovery can now be usefully accomplished and as scanning devices produce scans with more and more events, it will become more and more useful. However, as long as scans remain quantum limited, i.e. for the foreseeable future, processing routines for resolution recovery will need to vary from point to point as a function of the local intensity distribution.

In particular, variable filtering and density shifting are promising routines for resolution recovery. We plan to do more research on the use of these methods and on the development of these and perhaps better methods for resolution recovery.

Acknowledgements

The research reported in this paper is very much the work of a team. The author gratefully acknowledges the work of his colleagues, Mr. Henri Vetter of Biophysics Laboratory, New England Medical Center Hospitals, Dr. David Chesler of General Telephone and Electronics Laboratories Inc., Dr. Samuel Wilensky, Dr. Gordon Brownell, and Mrs. Linda Deveau all of Physics Research Laboratory, Massachusetts General Hospital. The research reported in this paper was supported by Atomic Energy Commission Contract No. AT (30-1)-3937.

REFERENCES

1. Nathan, R., *Digital Video-Data Handling,* Tech. Report #32-877, Jet Propulsion Lab., Pasadena, California, January, 1966.
2. Nagai, T., Fukuda, N. and Iinuma, T. A., "Computer-Focusing using an Appropriate Gaussian Function". *J. Nucl. Med.* **10,** (15): 209-212 (1969).
3. Chesler, D. A., *Resolution Enhancement by Variable Filtering,* Internal report (Physics Research Laboratory, Massachusetts General Hospital), January, 1969.
4. Pizer, S. M. and Vetter, H. G., "The Problem of Display in the Visualization of Radioisotope Distributions" *J. Nucl. Med.* **7,** (10): 773-780 (1966).
5. Pizer, S. M. and Vetter, H. G., "Processing Radioisotope Scans", *J. Nucl. Med.*, **10,** (4): 150-154, (1969).
6. Iinuma, T. A. and Nagai, T., "Image Restoration in Radioisotope Imaging Systems", *Phys. in Med. and Biol.,* **12,** (4): 501-509, (1967).
7. Vetter, H. G. and Pizer, S. M.: "A model for perception of quantum limited images and its application to radioisotope scanning." *Proc. 7th Intern. Conf. Med. Biol. Eng.* Stockholm, 1967, p. 113.

APPLICATION OF COMPUTERS TO TREATMENT PLAN RADIATION DOSE COMPUTATION

J. Garrett Holt

This age of computers has led to the development of a specialized technique to compute the radiation dose distributions within patients which I should like to call "Comprehensive Treatment Planning". Comprehensive treatment planning is designed to fulfill the requirements of each individual patient and takes into account the distortion of the radiation beam produced by tissue inhomogeneities and shape. The effect of surface curvature, lung, tissue, bony structure and external obstructions in the radiation beam on the dose distribution has been well established experimentally and techniques have been developed to compute dose distributions taking these parameters into account. The techniques used in comprehensive treatment planning are applicable primarily for high energy radiation and are based on Compton interaction, which shows only a slight dependence on the atomic number of the absorber. Technological development to produce isocentric high energy radiation therapy equipment made it possible to administer the radiation prescription more conveniently and with greater precision. As a result there has been an increased demand for more realistic treatment plans to yield an accurate and detailed representation of the radiation dose distribution within the patient. While it is perfectly feasible to perform these calculations manually, with the aid of isodose curves and a desk calculator, it is very time consuming and impractical to do on a routine basis for individual patients. Reference to an atlas of dose distribution, while useful in initial treatment planning consideration, is limited by the inaccuracies introduced by individual variations in patients structure.

Comprehensive treatment planning requires a great effort to accumulate accurate data on patients. Various techniques such as treatment simulation, horizontal transverse tomography, and density distribution scanning are employed to accumulate specific data in individual patients. To facilitate the rapid production of individualized treatment plans for each patient, computer-assisted methods of treatment planning have been developed.

By their ability to perform many computations rapidly and accurately digital computers, providing they are properly programmed, have proven to be an invaluable tool to the radiation therapist. The development of computer programs for radiation dose computation began about a decade ago and has been refined to a high degree of sophistication. The number of workers that have contributed to this rapidly growing field are too numerous to mention. The main effort has been directed towards providing dose computations for external beam therapy and interstitial and intracavitary therapy. The various approaches employed to compute the distribution of the radiation field within a medium may be grossly classified into three categories: fundamental, empirical, and curve fitting.

Perhaps the most fundamental approach is the computation of the radiation dose distribution by means of generating the history of a large number of single rays, utilizing Monte Carlo methods to account for the various interaction probabilities. This approach, while perhaps the most satisfying one intellectually, requires the calculation of many thousands of ray histories and is suitable only for research applications. A somewhat simpler method is the division of the radiation field into primary and secondary radiation and to compute each contribution separately. The method employed is based on Clarkson's method. The primary contribution to the absorbed dose is computed by an exponential computation while the scatter contribution is obtained by integrating the scatter functions or scatter air ratios over the radiation field.

The most commonly employed empirical method is the utilization of isodose data. In effect the calculations are quite similar to the ones which would be performed manually. The isodose data is stored in matrix form and the dose distributions are computed by matrix multiplication of the isodose matrix with a tissue air ratio column matrix. Another method of generating a dose distribution, employs tissue air ratios which are multiplied by the appropriate off-center-ratios (OCRs) to obtain the dose to points not located on the central axis. A similar method employs the so-called decrement line which is the line of constant OCR as a function of depth.

The curve fitting approach utilizes an analytical expression which has been made to fit the depth dose and OCR data. Instead of storing the depth dose and OCR information analytical expressions are used to compute the values as required.

The models employed for the calculation of dose distributions resulting from interstitial or intracavitary high energy gamma sources may also be classified into three categories: point source calculation, line source calculation, and empirical depth dose calculations. The first of these merely assumes a isotropic source and the computations involve a simple inverse square law calculation only. Linear sources under this approach are divided into a number of point sources. The second, in one form or another, utilizes the Sievert integral for filtered line sources and the integral is generally evaluated by a polynomial expansion. The third method, which is used for low energy emitters, utilizes a table lookup based on experimental depth dose curves.

Some programs may be used to compute the radiation dose distribution throughout any irregular 3-dimensional volume. This feature is particularly useful in the treatment planning for mantle fields employed in the treatment of Hodgkin's disease and in the computation for treatments requiring non-planar diverging beams, such as might be found in the treatment of head and neck disease. At this state of the art, however, 3-dimensional computations are impractical for large volumes. The cost of computing a 3-dimensional matrix may be of the order of several hundred dollars per treatment plan. In addition to this, it is necessary to have the patient's data available in 3-dimensions and the interpretation of 3-dimensional output is rather difficult also. While it is possible to store the 3-dimensional matrix and display arbitrary cuts through this matrix, by means of a suitable terminal, the cost of the hardware is prohibitive. Fortunately, the bulk of external beam radiation therapy is 2-dimensional which simplifies the problem considerably.

The empirical approach is perhaps the most commonly used method of computing the dose distribution within a plane. The method which employes isodose matrices is limited by the number of isodoses available and generally assumes a constant target-skin-distance, although an additional correction which corresponds to the Mayneord F factor may be used to correct for the change of depth dose with SSD. The TAR/OCR or percent depth dose/decrement line models are capable of generating isodoses for any rectangular field size. It has been shown that when tissue-air-ratios are plotted as a function of area/perimeter (A/P), rather than field size, it is possible to utilize this normalization to account for field elongation with an accuracy of about 2 percent. The OCR's are a function of depth and field size. However, the field size dependence is only slight and it is possible to describe all field sizes by 2 sets of OCR ratio tables. A combination of TARs and OCR ratios may therefore be utilized to compute planar dose distributions resulting from the application of rectangular fields of any dimensions. This feature minimizes the amount of information which must be stored and makes this approach very useful.

In all methods, except the fundamental Monte Carlo method, the presence of inhomogeneities is generally accounted for by expressing the radiation path length throughout the inhomogeneity in terms of equivalent unit density material. This assumption is valid only for high energy radiation and inhomogeneity corrections are applied only for the calculations for treatment planning with Co-60 or high energy x-ray beams.

For implant dosimetry, the critical volumes are usually small making it more reasonable to provide 3-dimensional calculations. The dose distribution throughout a volume is generally described by a series of dose matrix planes, which may be computed at any orientation.

The specific form of the input and output format depends on the available hardware but certain criteria must be met before a computer program is acceptable for clinical usage. As a minimum requirement computerized methods for dose calculations should be at least as accurate as hand computations. Any computer program for treatment planning must be capable to fulfill the various needs of the radiation therapist. The program should be sufficiently flexible to be used for all types of Co-60 teletherapy machines and high energy x-ray generators. Differences in source size, treatment distances, collimation, should be accounted for. The program should be capable of computing the dose distributions resulting from several co-planar fields of different size, as well as from wedge fields, rotating fields, arcing fields, diverging fields and weighted fields. In short, the program should accept any treatment situation which is likely to arise within the radiation therapy department. It is essential that the computations are based on proper corrections for patient's curvature and tissue inhomogeneities such as lung and bone. Failure to correct for the presence of lung tissue may lead to errors in dose of the order of 20-30%. The computed dose distributions must accurately depict the dose distributions within the patient; any computer program should be well tested before utilization on patients. The dose computations should be backed up by physical measurements in phantoms as well as *in vivo* measurements on patients. One word of caution is in order here. Realistic dose computations must be used with the full understanding and cooperation of the radiation therapist involved. Through years of experience the clinician has learned to expect certain clinical results following the traditional course of radiation therapy. If the radiation beam has penetrated inhomogeneities there is a strong possibility that the radiation dose as computed by traditional means may have been in error. If allowances are made in the computations for the presence of inhomogeneities, the resulting dose distributions must be fully understood and the possible clinical implications must be considered carefully.

Certain specific recommendations can be made to be used as a guide for the evaluation of a treatment planning computer program. Dose computations should be based on well established data, such as published TARs or central axis depth dose tables for Cobalt-60 radiation as found in BJR Supplement No. 10. In the case of high energy x-ray beams where no such published data are available, data used for computation should be carefully measured and published in the literature. Computer programs must accommodate the specific requirements of individual users. Source size and collimation greatly effect the general shape of the isodoses in the penumbra region. It should be possible to incorporate specific machine data such as penumbra information, wedge filter profiles, etc., since these parameters are likely to vary from machine to machine.

Input format should be as simple and non-technical as possible. A real time, direct dialogue between the radiologist and the computer would be ideal. While these systems are under development, it should be borne in mind that for a program to be generally useful the input format must be sufficiently simple that it can be understood and easily remembered by a non-computer oriented user.

Whatever type of output display is selected it should result in a hard copy, which may be in the form of a dose matrix, dose contour map, or isodose curve. While the dose matrix contains the maximum amount of information, isodose curves are easiest to interpret. The exact choice of graphics depends somewhat on the available equipment and the preference of the user. The use of automatic digitizers for tracing patients contours and the localization of sources from radiographic film is very convenient but rather costly. Video displays of dose information is also highly effective. While these systems are available now the cost is too high to justify their general usage.

A device known as the Programmed Console (P.C.) has been developed which utilizes a small computer to generate visual displays of isodose curves, provides a tremendous means of teaching treatment planning. The P.C. utilizes isodose curves which are digitized onto magnetic cards to generate dose distributions.

The system is entered by means of a keyboard or a positional transducer, which may be utilized to digitize a patient's contour or the single field isodose curves. The angle of entry of several planar fields of equal size may be manipulated by means of a manual control and the resulting isodoses within the patient's contour are rapidly generated by this system and displayed over the video screen. While an obvious compromise in accuracy was required in the calculations, the system provides a dramatic means of studying the effect of multiple fields on the dose distribution. For more accurate calculations the programs console may be interfaced with a larger computer.

Because of the high cost of computerization, routine treatment planning methods are utilized only by a few large centers, and the possibilities of this powerful tool in radiation therapy so far has not been widely utilized. The bulk of treatment planning in radiation, if it is done at all, is done manually for a few isolated cases. The need to extend the benefits of this powerful tool to institutions which lack the treatment planning and computer capabilities has long been recognized. In the past several institutions have provided treatment plan computations for other hospitals within their community. These treatment plan computations were either carried by courier, sent through the mail, or more recently, transmitted over the telephone lines by telecopier or teletypewriter. Some government support has been received for the development of computerized treatment planning centers. The general aim of computer centers is to promote better treatment planning in small institutions and to serve a large community by providing computerized treatment planning as a service. Part of a function of the computation service is to develop practical methods of telecommunication with collaborating hospitals. Another function is the development of education and consultation facilities for the community at large. This should include the development of instantaneous display systems of radiation dose distribution by means of video communication. Perhaps the most essential service a large center should be able to provide to a peripheral user is treatment planning consultation. The question most frequently asked is what is the best method of treatment for this lesion? In order to answer this question visual communication is nearly always required. In some cases the contour information may be sufficient information, but generally video or personal communication is preferred. Unfortunately, such systems as the picture-phone will not be generally available until about 1975. Another aspect of a consultation service is the establishment of standards of equipment calibration and tumor localization. It is of course essential that the treatment plan be faithfully reproduced in the patient.

The question for whom this service should be provided has not yet been settled. Should only full-time radiotherapists be eligible to receive treatment plans? Should anyone be refused? Opinions differ and time and experience will provide the answers.

The role of the radiological physicist in the development of computerized treatment planning is essential. He is most qualified to provide the link between the radiation therapist and the computer programmer. Since the methods of computerized treatment planning are still in the developmental stage, the development and evaluation of the necessary software should remain firmly the responsibility of the qualified physicist who is actively engaged in treatment planning. The establishment of treatment planning computation centers should be limited only to large institutions with adequate treatment planning facilities, a competent physics staff including full-time programming capabilities, and who are prepared to continue the development of existing programs to keep in step with the variety of needs of radiation therapy as well as the improvements in computer technology.

Perhaps the most exciting recent development is the utilization of commercial remote access computer services for treatment planning dose calculations. By having treatment planning computer programs accessible through commercial services, it is possible to disseminate the methodology rapidly on a nationwide basis. The location on-site of a large computer facility is no longer required and anyone who qualifies to do competent treatment planning may have access to computerized methods by subscribing to a commercial computer service. Another advantage of time-sharing is the rapid turn-around time and convenience. Since the utilization of commercial remote access batch or time-sharing services for computerized treatment planning is a very recent innovation, some control should be exercised over the commercial use of the treatment planning computer programs. The software should not be released completely to commercial interest since this would hinder their further development and with turnover of computer personnel within the companies would soon detach the programs from the programmers. Instead the authors should retain control over the computer programs and above all should retain the privilige of changing or modifying the programs to up-date the software as required. In addition to this, it would seem highly desirable that the institution from which the programs originated retain complete authority to select the potential users of treatment planning programs. This would assure that the programs are released only to qualified institutions. This kind of control may be accomplished by having the programs locked into the system in the object-code, so that they can be neither listed nor changed by the user. No user's number should be authorized until the potential user has contacted the author of the program and has been thoroughly briefed in the theory, assumptions and limitations inherent in the program. On the other hand, the originating institutions should be prepared to render all possible assistance to a user in the application of these programs. In our experience a minimum orientation session of one day in addition to about 25 supervised calculations is required before the user achieves a reasonable degree of competence in the use of the program. While these restrictions may appear somewhat severe to impose upon a commercial service, they are essential, at least initially, until adequate users guides and program documentations may be produced. At this writing at least 4 commercial services provide treatment planning dose distribution computations which are available by subscription to the service.

An example of such a treatment plan computation network, which is available through a commercial time-sharing service, is the Memorial Hospital Dose Distribution Computation Service. This service has been made available to qualified users on a national basis and has been in operation for about two years. It is supported by Memorial Hospital and is operated under the guidelines stated above. Users communicate directly with the commercial computers by means of a teletypewriter terminal and are billed monthly by Memorial Hospital for their usage of computer time. Fig. 1 depicts a model 33 teletypewriter with an acoustic silencer receiving a GYN radium dose computation. Figs. 2 and 3 show additional examples of dosage computations received over the teletypewriter and the treatment plans drawn from this information.

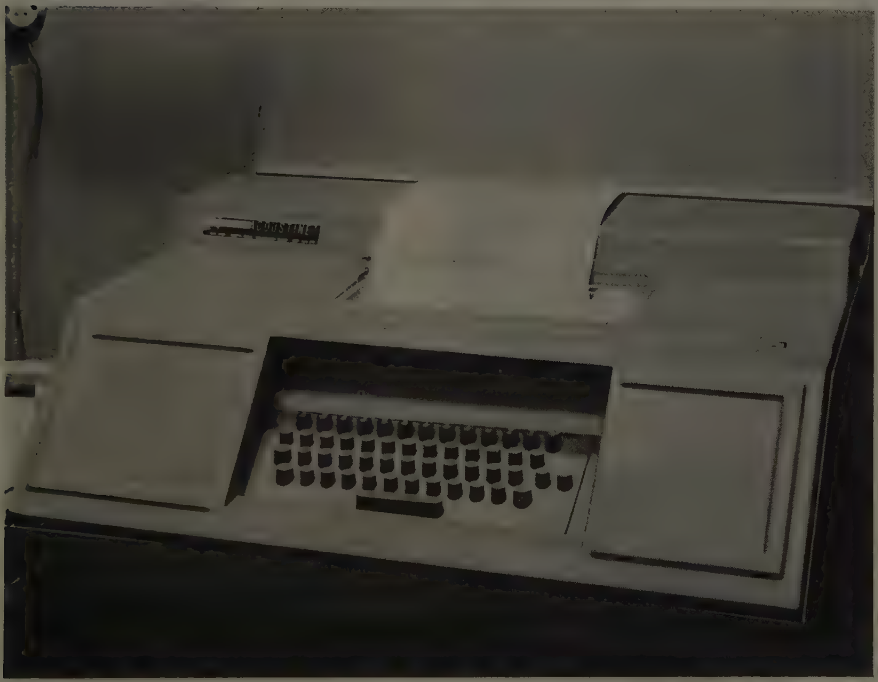


Fig. 1. Model 33 teletypewriter, which is used as a remote terminal to access a commercial time-sharing service, is shown receiving a GYN dose distribution calculation. With this terminal, a user communicates directly with the computer by means of ordinary telephone lines, and is responsible for coding his own input for treatment planning calculation.

After the successful development of computerized treatment plans which accurately depict the dose distribution in a medium, it is meaningful to reverse the question. Given a patient with a lesion in a certain location and given other areas within the patient which should be spared, what is the best way to treat this patient with radiation therapy? Indeed, an effort is being made to develop computerized methods of treatment plan optimization. Two methods of optimizing dose distributions are presently under investigation. One method assigns a figure of merit to selected target areas within the contour and computes an optimum score by means of a series of vector dose calculations. The other method divides the plane of the radiation field into 72 dose vectors. Within the treatment region upper

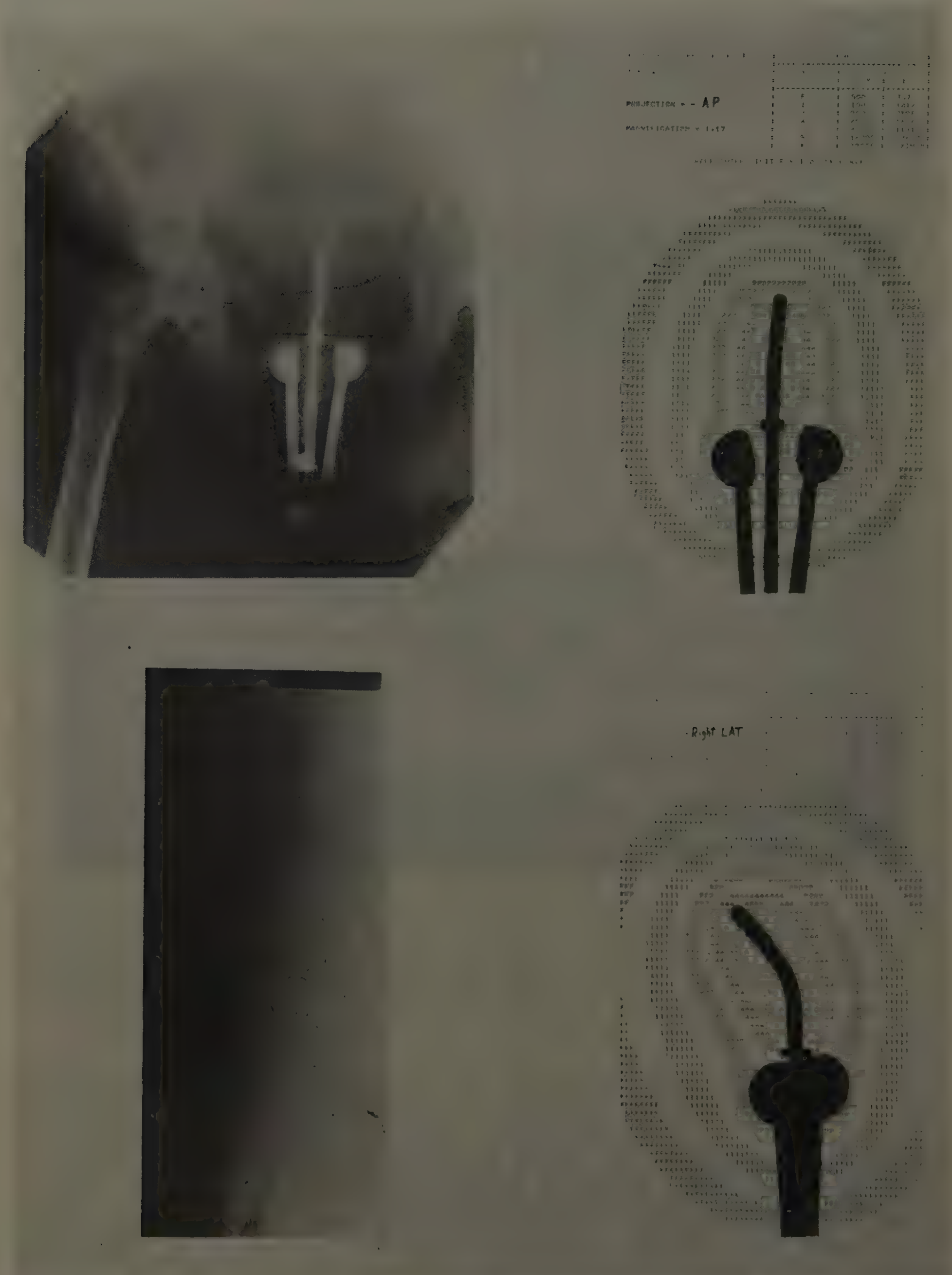

PROJECTION - - AP
MAGNIFICATION = 1.17
Right LAT

Fig. 2. Dosage calculations for gynecological radium application. The tandem contains 3 radium tubes: 25 mg., 12.5 mg. and 12.5 mg. and the ovoids contain 12.5 mg. of radium each. The dosage distributions indicate rads/3500 mg-hrs.

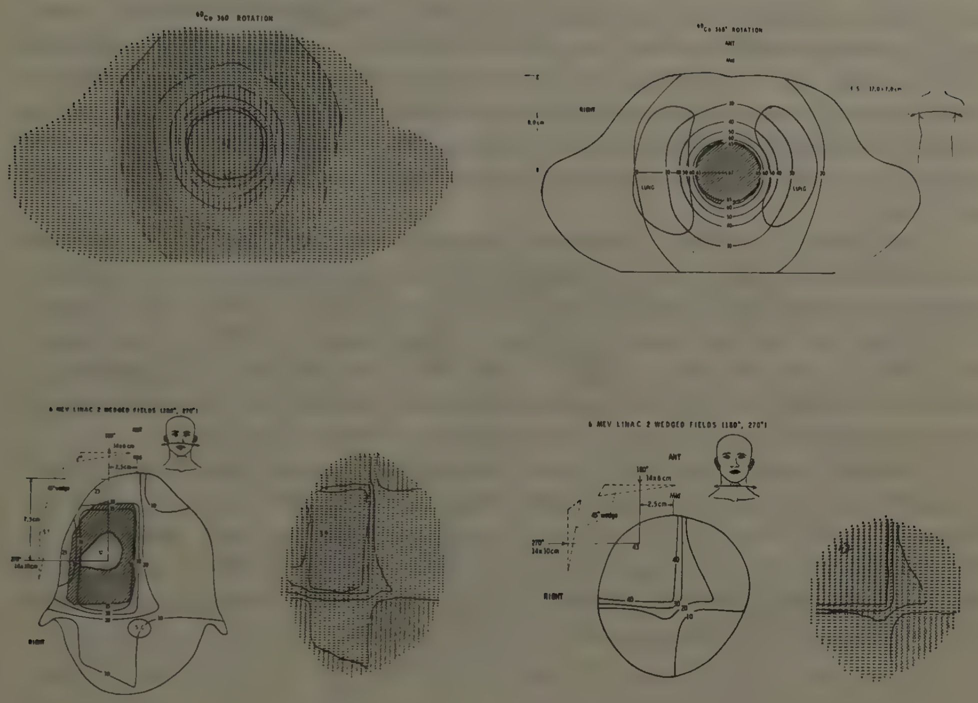

⁶⁰Co 360 ROTATION
⁶⁰Co 360° ROTATION
ANT
Mid
RIGHT
8.0cm
T.S. 17.0×7.0cm
LUNG
LUNG
6 MEV LINAC 2 WEDGED FIELDS (180°, 270°)
2.5cm
7.3cm
45° wedge
S.C.
RIGHT
6 MEV LINAC 2 WEDGED FIELDS (180°, 270°)
ANT
180°
14×8cm
Mid
2.5cm
45° wedge
270°
14×10cm
RIGHT
40
30
20
10

Fig. 3. The upper part shows dosage calculations for a 360° rotation with Cobalt-60. Patient's curvature and the presence of the lungs in the radiation field have been considered in the calculations. On the left hand side is the dose distribution map as received over the teletypewriter, and on the right hand side is the finished treatment plan as drawn from the dose contour map. The bottom shows the dose distributions for two wedged fields with 6 MeV x-ray radiation. The treatment plan has been drawn in the two planes indicated.

and lower bounds may be assigned to certain sub-regions and the best strategy of the weighted field vector distribution is computed be methods of ordinary linear programming. While these methods of optimization are not designed to replace the radiation therapist, the solution to this problem would serve to minimize the amount of guesswork and intuition which is still present in the initial selection of the radiation prescription.

One important problem which still needs to be solved is an evaluation of the various computer programs which are in current use. It is an unnecessary duplication of effort for anyone who wishes to get into the field to duplicate the years of programming required to produce a functioning treatment planning computer program. It is more logical to utilize one of the existing programs which have been established already. However, at this time very little information is known regarding the accuracy of the various programs and there is little to guide the potential user as to which program is more suitable to his needs. Such professional or scientific organizations as the HPA and AAPM would be well qualified to establish standards for computerized treatment planning. Although there have been individual efforts to demonstrate the accuracy of various programs by comparing computer generated isodose curves, no overall evaluation program has been developed so far. While the formulation of a set of standards for computerized treatment planning may be accomplished without too much difficulty, the evaluation of the numerous programs now in use is a formidable task which will require a great deal of time and effort. However, steps are being taken at this time to investigate these problems.

The cost of computerized treatment planning varies considerably and depends largely on the facilities available. Utilizing a commercial service one could estimate the cost of an interstitial or intracavitary source calculation to be of the order of $10-$15 and an external beam calculation to range from $20-$100, depending on the complexity of the plan. In view of this, the establishment of a "dosimetry" charge to the patient would not seem unreasonable. In any case, it would seem a small price to pay compared with the overall cost of radiation therapy.

Treatment planning radiation dose computation is merely one area in which computers are being used in Medical Physics. Computers have provided an extremely powerful tool to the medical physicist and the potential application of computers in medical physics so far has been barely scratched. The next decade is bound to produce sweeping changes in the computer hardware which should increase the use of computers in para-medical applications.

OPTICAL IMAGE PROCESSING

Brian J. Thompson

INTRODUCTION

Optical processing of data is a title which covers a very broad spectrum of technologies. Several volumes have been devoted to the subject. (1,2) Clearly then in this paper the total field is not going to be covered. However I hope to make some pertinent remarks on the role that optical processing could play in medical physics and provide a follow-up to an earlier paper on a related topic. (3)

The wide variety of techniques and technologies that have potential for optical processing leave the uninvolved somewhat dismayed. However a number of methods are already being exploited. The most widely known is the use of two dimensional scanning of photographic data to produce a quantized plot of the density variations occurring on the film. Figure 1 shows typical results obtained with one type of instrument called an isodensitracer (4-6). Essentially a contour plot is produced with the intervals between the elements of the three symbol code (space, dot, dash for increasing density) being as small as 0.01 density units. I will return to some comments on scanning devices later.

The main topic of this paper is a discussion of some relevant techniques in coherent optical processing.

COHERENT OPTICAL PROCESSING

The idea of using coherent optical systems in such a way as to allow for the manipulation of optical images is not new. The basic ideas are included in Abbe's (7) theory of vision in a microscope first published in 1893. More specifically the illustrative experiments of Porter (8) (1906) are certainly prophetic of much of present day optical processing. To interpret Abbe's ideas somewhat loosely it is true to say that he realized that image formation in a microscope was more

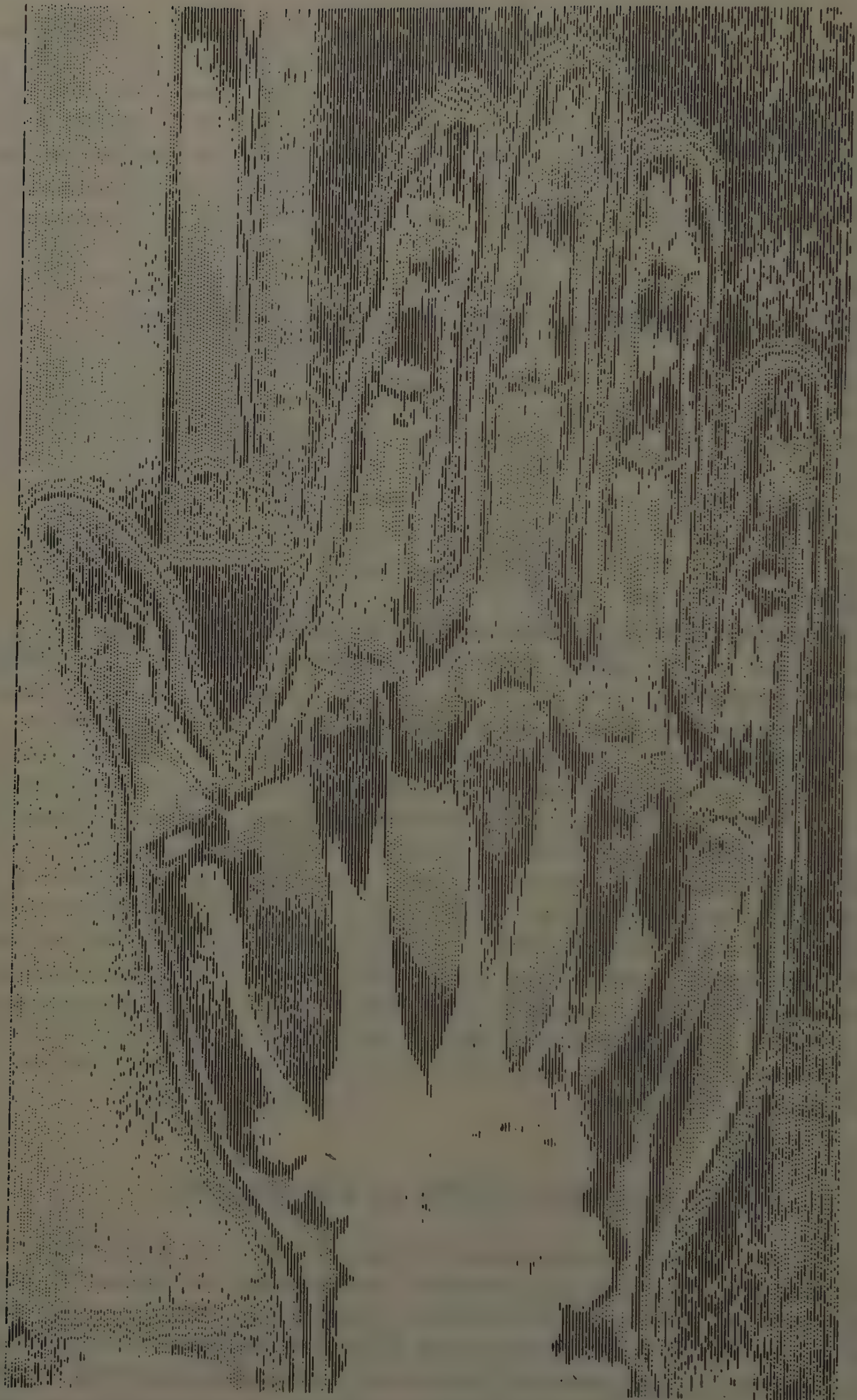


Fig. 1(a). Isodensitometer traces of x-ray record of a hand, density increment = 0.08 and magnification 1×.

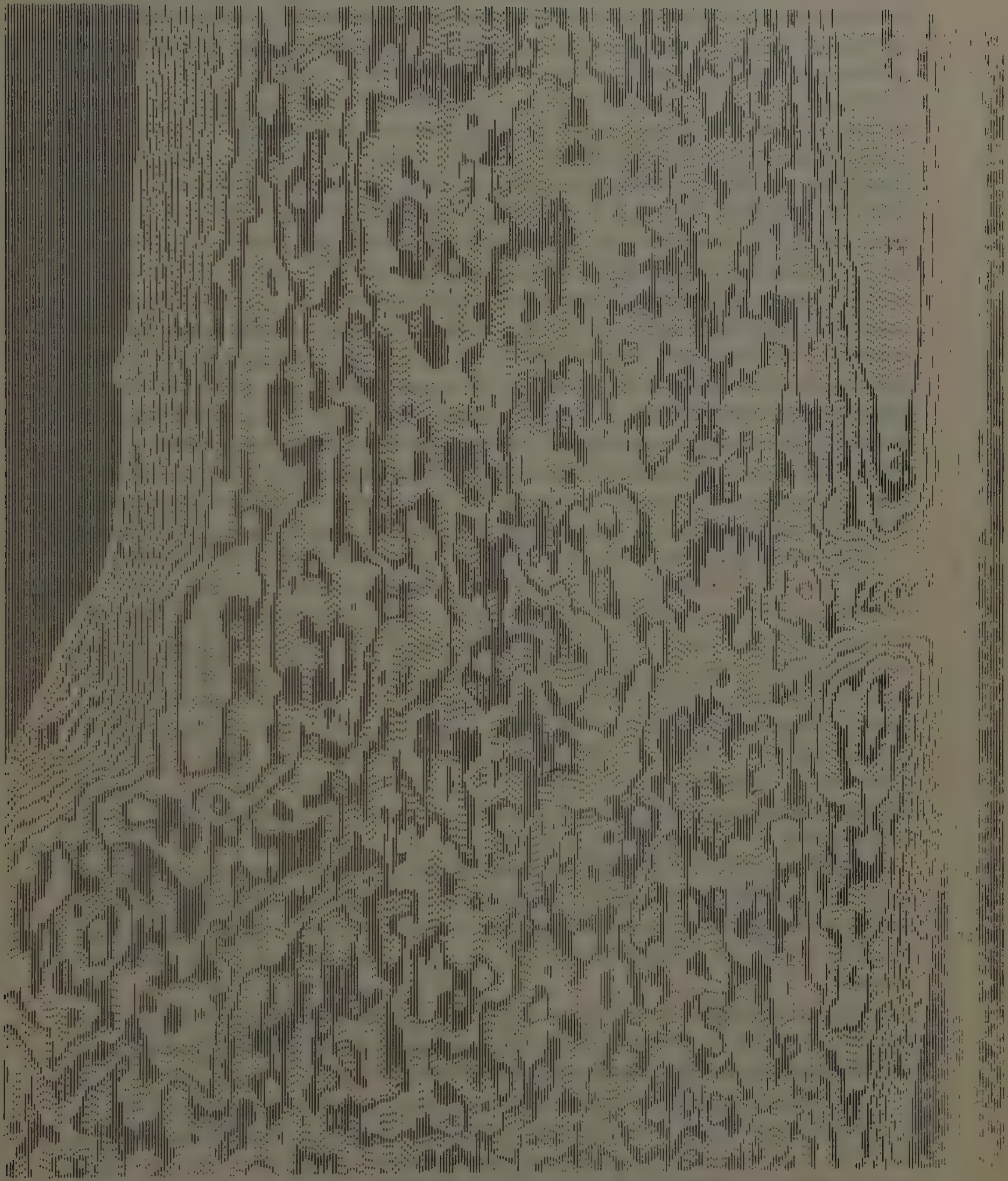


Fig. 1(b) Isodensitometer traces of X-ray record of an ankle, density increment 0.02 and magnification 10×.

correctly described as a coherent process rather than the more familiar incoherent process. Abbe theorized that the light illuminating the object of interest in a microscope would be diffracted (a coherent light phenomenon). To form an image clearly this diffracted light must be collected by the objective lens and the nature of the image and the resolution would be affected by how much of that diffracted light was collected (by comparison we may recall that in incoherent image formation the nature of the image is not changed; only the resolution is affected by the change in numerical aperture of the imaging system). By way of an example we can consider an object consisting of a periodic variation in amplitude transmittance. The light diffracted by this object will exist in a series of discrete directions (or order of diffraction). This series of orders contains a zero order propagating along the optical axis and a symmetrical set of orders on either side of this zero order. Abbe correctly speculated on what would happen as the microscope objective accepted different combinations of these orders. What information is carried by the zero order alone? What is the nature of the image as other orders are included? For instance if the zero order and one first order are collected then the image information will be that the object consisted of a periodic distribution and that the correct periodicity is determined. However the spatial location of the periodic structure is not correctly ascertained. If now the other first order diffracted light is included, then the correct spatial location of the periodic structure is also included. Additionally, however, a small spurious intensity peak will become visible between the main intensity peaks defining the periodic structure. The nature of the image is dependent upon the diffracted light collected. As more orders are included the image more closely resembles the object. We note the condition that for a minimum acceptable image we must include the zero order and the two first order diffracted beams. We must comment again that by comparison the incoherent image under similar circumstances will not show the type of spurious effects discussed above. Finally we may examine the nature of the image if the first orders are collected but the zero order is omitted. The image appears to have twice the frequency of the original object! Abbe's theory of vision in a microscope was a description of image formation in terms of the addition of interfering beams—these beams being the diffracted light from the object.

Porter showed that these ideas were correct in an excellent series of experiments to determine the effects on the image of a periodic object by removing various combinations of orders of diffraction. The Porter experiments are excellent examples of simple blocking filters and are often repeated today for illustrative purposes.

These two pieces of work were really the starting point but they are also the forerunners of the process that is now called spatial filtering. Other examples are Schlieren systems and the phase contrast microscope.

Schlieren systems have been known and used in many forms for a variety of purposes associated with the display of index variations caused by gas flows etc. This type of object is, of course, purely a phase object since its only effect is to produce a difference in optical path. Schlieren systems are related to the Foucault knife edge test that was designed for the study of phase defects in lenses and mirrors. These systems clearly had an influence on the invention of the phase contrast microscope by Zernike (9) which is also an example of spatial filtering. The phase contrast microscope is applicable to the display of small phase differences in the object as an intensity distribution in image space.

Basic Optical System

The basic system required for coherent optical processing (or optical spatial filtering) is that shown as a two lens imaging system in figure 2(a). A collimated beam of quasi-monochromatic coherent light is used to trans-illuminate the object. Originally the illumination had to be derived from an arc source filtered for one of the prominent lines and then imaged onto a small aperture and the light allowed to propagate. Today of course the use of the laser is an attractive alternative. The first lens produces an exact Fourier transform of the object amplitude transmittance in its rear focal plane. We recall that this Fourier transform is just the Fraunhofer diffraction pattern associated with the object. The second lens then retransforms this transform to an image. Again the image amplitude distribution is equivalent to the object amplitude with no phase errors and the image lies on a plane. Purely from geometrical optical considerations this system is easily described as a simple imaging device that produces a magnification in the ratio of the focal lengths. If we depart from the exact location of the object then of course

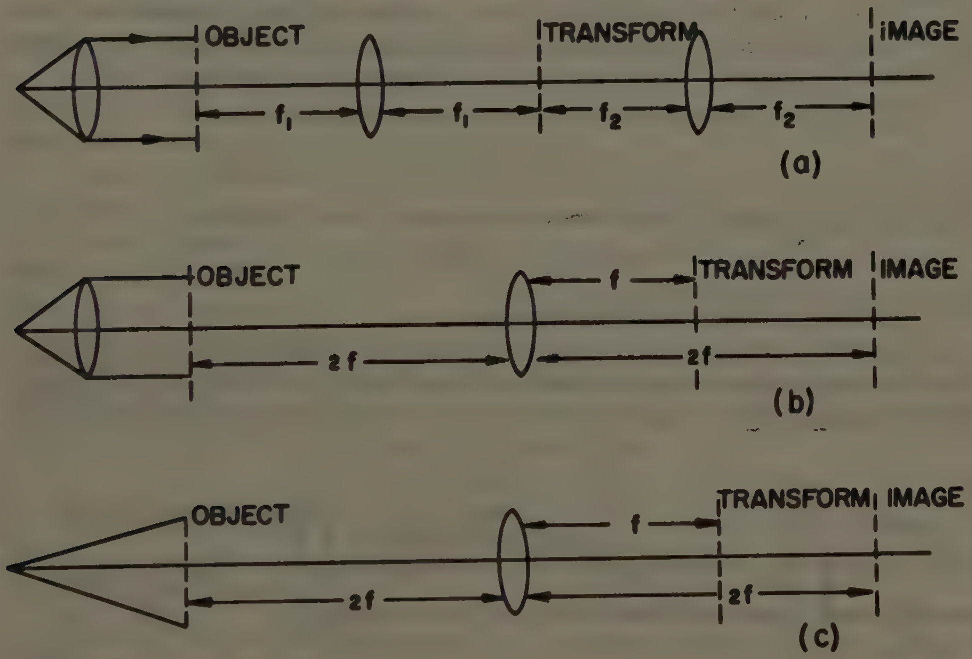


Fig. 2. Schematics of typical coherent optical systems for data processing.

an image of the same magnification is still formed in a different place and the image amplitude has associated with it a quadratic phase factor. The original systems of this type for image processing were not in this exact configuration however the two lens systems were used. This experimental arrangement Maréchal called a "double diffraction" arrangement for obvious reasons. Now it is simply called an optical spatial filtering system. The word spatial filtering arose from the operation of the optical device. In the intermediate plane the spatial frequency spectrum associated with the object is displayed. This spectrum may then be operated upon by a filter and the nature of the image changed in a controlled way.

Figures 2(b) and (c) show two other alternate systems that can be useful when phase considerations are no longer important. In (b) the transform plane is in the focal plane of the lens but this same lens is also used to form an image of the object. In the third system one lens does the complete job; the object is illuminated with a diverging beam and its transform is formed in the image plane of the source.

System Analysis

To examine the process in a little more detail we will use the system outline in figure 2(a) and define the amplitude transmittance of the object as $A(x,y)$, the distribution in the transform plane as $U(\xi,\eta)$ and the final image amplitude as $\psi(x',y')$. The coordinate system is indicated in figure 3. The coordinates $x'y'$ are in a reflected geometry to avoid the negative signs that occur because of the double transformation process. (The image is, of course, inverted in that plane.)

The first lens produces an exact Fourier transform and hence

$$U(\xi,\eta) = C_1 \iint A(x,y) \exp - ik(x\xi + y\eta)/f_1 \, dxdy, \tag{1}$$

where $k = 2\pi/\lambda$ and λ is the mean wavelength of the illumination.

$$U(\xi,\eta) = C_1 \tilde{A}(\xi/\lambda f_1, \eta/\lambda f_1); \tag{2}$$

the tilde indicating the Fourier transformation.

A general operation is then performed on $\tilde{A}$ to achieve some desired result. This filtering operation we will denote by $F(\xi,\eta)$, where this filter may be a single opaque blocking filter or may operate on either the amplitude and/or phase of the

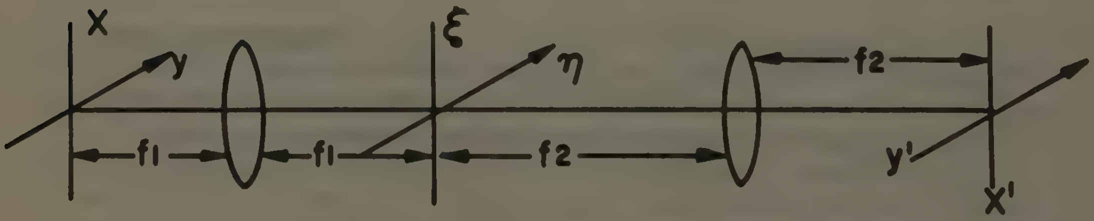


Fig. 3. The coordinate system.

light distribution. The complex amplitude distribution immediately after the filter $U'(\xi,\eta)$ is given by

$$U'(\xi,\eta) = U(\xi,\eta)\, F(\xi,\eta) = C_1\, \tilde{A}(\xi/\lambda f_1,\, \eta/\lambda f_1)\, F(\xi,\eta). \tag{3}$$

The second lens now forms an image by transformation of U'. Hence

$$\begin{aligned}\Psi(x',y') &= C_1C_2 \iint \tilde{A}(\xi/\lambda f_1,\, \eta/\lambda f_1)\, F(\xi,\eta) e \frac{-ik}{f_2}(x'\xi + y'\eta) d\xi d\eta \\ &= C_2C_2\, A\left(\frac{x'f_2}{f_1}, \frac{y'f_2}{f_1}\right) * \tilde{F}\left(\frac{x'}{\lambda f_2}, \frac{y'}{\lambda f_2}\right), \end{aligned} \tag{4}$$

where the * means a convolution. Note that the magnification factor f_2/f_1 arises naturally. The resultant intensity distribution in the image is $\Psi(x',y')\Psi^*(x',y')$, the star here meaning complex conjugate.

Filter Realization

The filters can be grouped into a number of types.

Blocking filters: Here certain spatial frequencies or bands of spatial frequencies are removed completely. These filters are simple to make and use but problems can occur because of their sharp cut-off which can produce "ringing" in the same way that a sharp cut-off filter in an electrical circuit produces ringing.

Amplitude filters: These filters operate in a continuous way on the amplitude (but not the phase) of the transform. The variation in amplitude transmittance required is often produced by recording the appropriate distribution on film.

Phase filters: Here only the phase is operated upon as in some phase contrast microscopes. Phase distributions are difficult to manufacture in a controlled way especially if continuously varying phase is required.

Phase and amplitude filters: In this situation the amplitude and phase of the optical field is operated upon. Clearly this is the most general type of filter and is the most difficult to construct. Orginally the amplitude and phase portions were made separately with the amplitude being made as indicated above. The phase filter can be manufactured by using an evaporated layer of material such as magnesium fluoride. Today the most efficient technique for fabrication of the filter is by an interferometric method in which the required complex amplitude is recorded as a hologram. Very often the required filter can be generated by the system without specific knowledge of the specifications.

APPLICATIONS

Simple Blocking Filters

A number of useful applications exist for very simple blocking filters. Included amongst these are the removal of rasters and half-tones, separation of additive noise, signal separation by orientation, etc.

Raster removal: A popular and sometimes useful application of a filter is to remove a raster from a television type picture. We can consider the input then as a periodic function whose envelope function is the continuous scene. Hence the transform consists of a periodic distribution in one direction with a periodicity reciprocally related to the object periodicity. Centered at each of these locations is the transform of the envelope (i.e. the scene spectrum). The individual members of these replicated spectra do not overlap. Hence if the filter selects one single order and the area surrounding it, the information about the periodicity will be lost and the scene information retained yielding a continuous tone image without a raster. This argument is readily extended to the two-dimensional case of a half-tone.

The question naturally occurs, why remove the half-tone at all since no extra information is added? Cosmetically the image is more pleasing without the half-tone. However there is a more important reason. When looking at the image under high magnification so that the field of view only contains a few half-tone dots (say 10 x 10) it becomes very difficult to "see the image" and hence interpret what is there. We can't see the image for the half-tone!

Removal of Additive Noise: If we have an object to which is added a noise contribution then the spatial frequency spectrum consists of an addition of the spectrum of the object and the spectrum of the noise. It is sometimes possible to partially separate out the object from the noise because the object spectrum may exist in known locations in the transform plane. For example if the object consisted of a periodic distribution such as a grating then its spectrum is located in a discrete set of orders. On the other hand, the noise spectrum is approximately some Gaussian distribution. If now a filter is used that lets through the discrete orders from the periodic object and blocks the rest of the spectrum then the periodic object will be enhanced with respect to the noise. It is important to note this operation quite precisely: we enhanced a known signal in the presence of an additive noise.

Similar results can be obtained with an isolated signal in the presence of additive noise. It is necessary to know what the signal is but not its location in the object space. Remember that the spatial frequency spectrum of the object is centered on the optical axis independent of its x-y location in object space.

This is an interesting result. However it has led to a great deal of confusion. Several workers have attempted to use this technique for the removal of so-called "grain noise" in photographic images. This is quite a different problem. Firstly the noise is not noise in the usual sense since the grains are the image; it is then clearly signal dependent. Secondly the grain is a multiplicative function in object space, i.e. the object is multiplied by the grain. Hence in the transform the object spectrum and the grain spectrum are convolved. A convolution is much harder to undo than is mere addition. The presence of the multiplicative grain is the cause of great difficulties in optical processing of photographically recorded images.

Signal Separation by Orientation: Very often in optical processing the signal that is required to be enhanced has to be known. Sometimes, however, it is equally valuable to know something about the noise or unwanted portion of the

Fig. 4. Optical filtering by orientation (After Beckerer and Thompson).

object. An example of this which illustrates also the importance of having a two-dimensional spectrum is that of selecting a signal from a set of seismic type traces. Figure 4(a) shows a set of 25 pen recorder traces, one of which contains a signal of importance. It will be noted that the quiet traces are essentially information in the horizontal direction whereas when a signal is present the pen trace contains a predominance of vertical lines. Hence a simple filtering operation removes all the spatial frequency spectrum in one direction (that corresponding to the quiet trace) and retains the remainder of the spectrum. The filter is then simply an opaque strip. The processed image is now shown in figure 4(b). Clearly the signal remains—also the timing marks. This was accomplished without knowledge of the signal or its position in the field of view. This type of filtering may well have application to electrocardiograph records.

Amplitude Filters

Contrast-enhancement: In the transform of a low contrast scene the zero-order diffraction term (or D.C. term) is large and the higher frequencies contain relatively little energy. If a small amplitude filter is placed to cover the zero-order to allow less of this light to pass through to the image plane, the other frequencies are not changed. Hence we have changed the relative energy distribution in the spatial frequency spectrum. The image now has improved contrast, the improvement depending upon the amplitude transmittance of the filter.

Differentiation—Contrast Reversal: The type of amplitude filter described above could be used in a more general sense to change the balance of the high and low frequency contributions to the image. Since edges in an object represent high frequency information it is clear that the edges of the object may be enhanced. Figure 5(a) shows a high contrast bar target that comprised the original object. A filter that has an amplitude transmittance that has a maximum in the center and then tapers linearly to zero at an intermediate frequency value is used to operate on the spatial frequency spectrum of the bar target object: the effect on the larger bars is to differentiate, i.e. only the edges are visible—but as bright

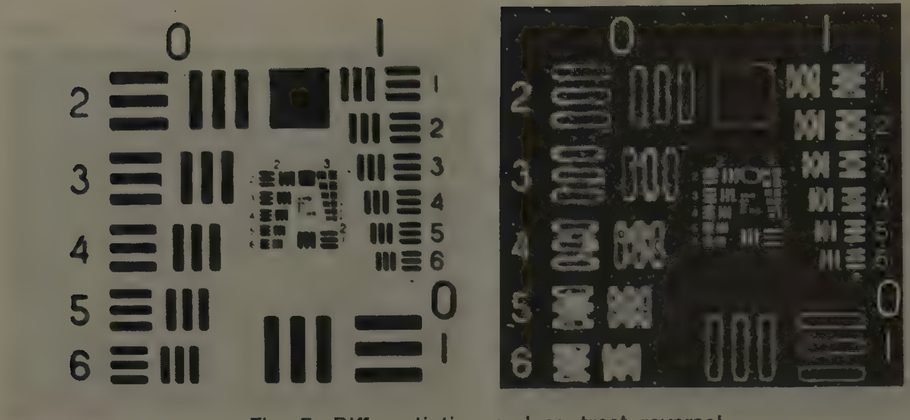


Fig. 5. Differentiation and contrast reversal.

outlines on an otherwise dark background. At higher frequencies a contrast reversal has taken place. The bars now appear as bright on a dark background. Notice, however, that this reversal is a one-way process. It isn't possible to start with figure 5(b) and obtain figure 5(a) by the same technique.

Phase and Amplitude Filters

Original work using complex filters was applied to phase contrast microscopy and aberration balancing. For excellent reviews see Tsujiuchi (10) and Goodman. (11) The great difficulties with the early efforts was the fabrication of the filter itself, however satisfactory results were obtained for image motion compensation, defocus correction and astigmatism. In recent years holographic techniques for the fabrication of filters has allowed these earlier methods to be implemented more easily and has led to considerable extension of these ideas. (12-16)

Again in this section instead of describing in detail the fairly numerous results I will pick on one particular result that may well have application.

Holographic Subtraction: In this method (16) it is possible to form a hologram filter from a particular object image and then use this filter to process other images of this same object so that an image of the difference is observed. Some of the detail of this method can be found in reference 16. In principle however, the optical system of figure 2 may be used with the modification that a second, reference beam is added to illuminate the transform plane so that a hologram may be formed there. A hologram is made of the first object. This hologram is processed as a negative and repositioned in the plane in which it was recorded. The original input is now replaced by a second object. The hologram filter now removes the first object information but allows the remainder to go on to form an image.

To illustrate the principle of this method an object consisting of seventeen circular holes, 1 mm in diameter, were drilled in a piece of thin brass to form a cross. All but two were then masked off. These two holes constitute the first object and as such were used in making the Fourier transform hologram.

Figure 6(a) is a photograph of the output plane showing the Fourier transform of the light distribution emerging from the hologram plane when the hologram is illuminated with the reference beam alone. The remaining holes were then unmasked thereby allowing the entire array to constitute the second object. Figure 6(b) is a photograph of the output plane when the hologram is illuminated with the Fourier transform of the second object. These photographs give only intensity distributions and therefore do not show that the light beams which formed these two patterns were out of phase by 180 deg, but this is nicely demonstrated by figure 6(c), a photograph of the intensity distribution in the output plane when both the reference beam and the second object beam illuminated the hologram. Note that figure 6(c) is clearly the difference between figure 6(a) and figure 6(b).

While this method has been discussed in terms of the familiar Fourier transform technique, it can obviously be generalized. Whether the optimum plane for

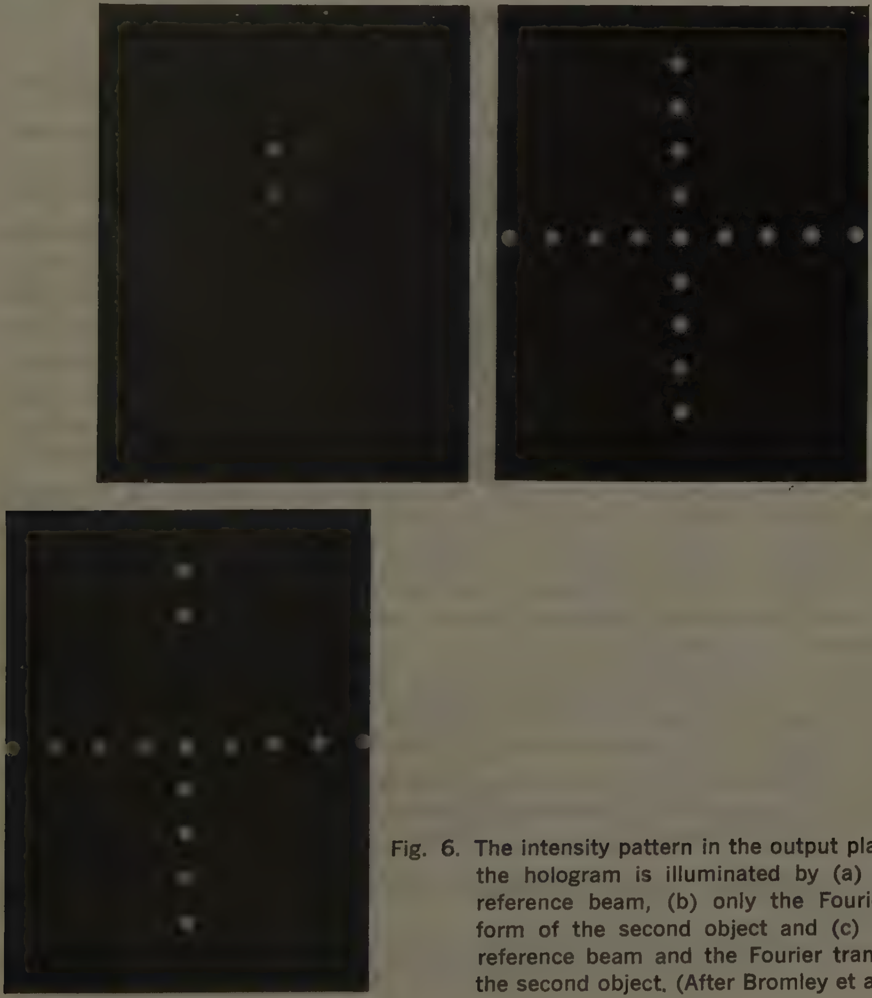


Fig. 6. The intensity pattern in the output plane when the hologram is illuminated by (a) only the reference beam, (b) only the Fourier transform of the second object and (c) both the reference beam and the Fourier transform of the second object. (After Bromley et al.(16))

the hologram recording is the Fourier transform plane or some other appropriate plane in the system depends upon the nature of the input object.

CONCLUSIONS

The impact of a combination of the principle of optical data processing and the newer techniques of holography will very soon be felt in a variety of applications in medical science and technology. Indeed recent work is already starting to produce results. (17, 18)

These efforts will be supported and augmented by computer techniques. A combination of optical techniques for presorting of data and then computer techniques for final analysis of the sorted data appears to be attractive.

REFERENCES

1. *"Optical Processing of Information,"* Eds. D. K. Polack, C. J. Koester, and J. T. Tippett, Spartan Books Inc., 1963.
2. *"Optical and Electro-optical Information Processing,"* Ed. J. T. Tippett, D. A. Berkowitz, L. C. Clapp, C. J. Koester, A. Vanderburgh, M.I.T. Press, Camb., Mass. 1965.
3. Thompson, B. J. *"Optical Processing of Visual Images,"* Invited paper AAPM Symp. on Diagnostic Radiology, San Francisco Medical Center, June, 1968.
4. C. S. Miller, F. G. Parsons and I. L. Kofsky, "Simplified two-dimensional microdensitometry," *Nature* **202,** 1196, 1964.
5. I. L. Kofsky, *"Reduction of Pictorial Data by Microdensitometry,"* Proc. Symp. Aerospace Measurement Techniques, M.I.T., July 7-8, 1966.
6. C. S. Miller, *"Isodensitometry as an aid in the analysis of photographically recorded information,"* 101st Technical Conference of S.M.P.T.E., April, 1967.
7. E. Abbe, Arch. Mikroscop. *Anat.* **9,** 413, 1893.
8. A. B. Porter, "On the Diffraction Theory of Microscope Vision," *Phil. Mag.* **11,** 154, 1906.
9. F. Zernike, "Das Phasenkontrastverfahren bei der Mikroskopischen Beobachtung," *Z. Tech. Phys.* **16,** 454, 1935.
10. J. Tsujiuchi, "Correction of optical images by compensation of aberrations and by spatial frequency filtering," *Progress in Optics,* **11,** 133, 1963.
11. J. Goodman, *"Introduction to Fourier Optics,"* McGraw-Hill 1969.
12. E. N. Leith, J. Upatnieks and A. VanderLugt, "Hologram microscopy and lens aberration compensation by use of holograms," *J. Opt. Soc. Am.* **55,** 595, 1965.
13. A. VanderLugt, *"Signal Detection by Complex Spatial Filtering,"* IEEE Trans. on Information Theory, **10,** 139, 1963.
14. G. Stroke, "A new holographic method for an a posteriori image-deblurring restoration of ordinary photographs using extended-source lensless Fourier transform holography compensation," *Phys. Letters,* **27A,** No. 7, 405, 1968.
15. D. J. Raso, "Simplified method to make hologram filters for target recognition," *J. Opt. Soc. Am.,* **58,** 432, 1968.
16. K. Bromley, M. A. Monahan, J. F. Bryant and B. J. Thompson, "Complex Spatial Filtering by Holographic Fourier Subtraction," *Applied Physics Letters,* **14,** No. 2, 67, 1969.
17. E. J. Feleppa, "Biomedical applications of holography," *Physics Today,* **22,** No. 7, 25, 1969.
18. *"Making photographs clearer,"* Research Trends, Cornell Aeronautical Labs., Winter 1969, p. 10.

GENERALIZED GRAYNESS PROFILES AS APPLIED TO EDGE DETECTION AND THE ORGANIZATION OF CHROMOSOME IMAGES

M. L. Mendelsohn*, B. H. Mayall, and B. H. Perry

The delineation of natural boundaries is an important early step in the analysis of many types of images. When the boundaries are sharp, and in the extreme case when the image is black and white, the problem of boundary location may be trivial. But when edges are gray and blurred, and when measures based on spatial dimensions of the object are important to the analysis, the definition of what is meant by a boundary and the precise location of it once it is defined can be a difficult and crucial problem. This is particularly true of many microscopic applications in which the objects may be inherently fuzzy, and the image is further degraded by optical diffraction effects and aberrations in the imaging system. In digital images these considerations are compounded by the potential degradation associated with spatial sampling and digitization of grayness.

The image of a metaphase cell contains many more or less discrete chromosomes each of which is potentially identifiable on the basis of morphology. In the digital analysis of such images the boundary problem takes two forms. At the outset the boundary is used to partition the metaphase image into objects corresponding to the individual chromosomes. Later the boundary is again required to estimate parameters for each chromosome, including area, length, stain content, degree of compaction, and location of the centromere.

In previous publications (5, 6, 10) we have discussed the use of histograms of optical density for the initial setting of overall thresholds in metaphase and cell images. Such methods work quite well in the partitioning of metaphases. But our early results indicate that individual chromosomes vary considerably in their darkness, their sharpness, and their surround. Therefore an overall threshold is seldom optimal for individual chromosomes and it is not adequate for the later and more finely detailed phases of the analysis. This same limitation of overall threshold also

---

* This work was supported by USPHS Contract PH 43-62-432, and USPHS Grants 5 K06 CA18540 and GM-16,913, National Institutes of Health, Bethesda, Maryland.

is applicable to images in which analogue or photographically imposed cut-offs are used in place of digital thresholds.

In this paper we describe an approach to setting individual chromosome thresholds. The method is based on an image transformation in which grayness or optical density is expressed as a function of distance from a provisional boundary. In addition to its intended purpose, the transformation gives some insight into the structure of chromosome images and it may have a broad utility in the analysis of gray images in general.

Chromosome Boundaries as Imaged by the CYDAC System

The CYDAC System produces digital images consisting of matrices of 200 by 200 eight-bit measurements of optical density. It images through a flying spot microscope operating in the blue-green region of the visible spectrum (5). In this instrument, as in scanning microscope systems in general, the density or grayness profile of a sharp high-contrast edge is a function of 1) the wavelength and coherence of the illumination, 2) the effective numerical aperture of the objective, 3) the quality of the optics including focus and lens aberrations, and 4) the size of the scanning spot source. This behavior can be predicted to a certain extent from theory (2, 4, 8, 11, 13), and it can be measured directly using appropriate test objects. Figure 1 shows the line spread function of CYDAC derived from an analogue scan across an edge of a metallic film imaged under optical conditions comparable to those used for chromosomes. The corresponding image of a biological edge would be the convolution of this line spread function and the true grayness profile of the object.

Although chromosomes are highly organized objects, they are not bounded by a membrane and they owe their organization to an internal rather than an external structure. In electron microscope images, such as DuPraw's (3), chromosomes consist of skeins of string-like fibers with haphazard edges and disorganized loops of fiber extending out from the main conglomeration. A correspondingly ragged appearance of the edge is sometimes seen in the light microscope, but in general the fibrous detail shown by electron microscopy is well below the limit of resolution of light microscopy. From the pattern of chromatid fibers in the electron micrographs one would predict that the grayness profile transmitted through the light microscope would be unsharp along the entire length of the chromatid.

The actual appearance of the grayness profile of CYDAC-generated chromatid images can be approximated from a series of scans taken perpendicularly to isolated chromatids. In the example to be shown such chromatids were found in the midst of a standard metaphase image. Appropriate segments of vertical or horizontal scan lines were selected on the basis of lack of interfering material and a corresponding horizontal or vertical orientation of the chromatid. Figure 2 shows a composite of 27 such chromatid profiles, all from the same scan, and all aligned so that their centers fall at zero on the horizontal scale. The general appearance of this composite is a symmetrical bell. Composites from other metaphases give similar shapes, except that metaphases with dark highly compacted chromatids give broader and higher bells.

There is a striking difference in sharpness displayed by Figure 1 and Figure 2, leading to the conclusion that diffraction effects are relatively minor in Figure 2 and hence that the predominant factor in the grayness profile of the chromosome image must be the inherent structure of the chromosome itself.

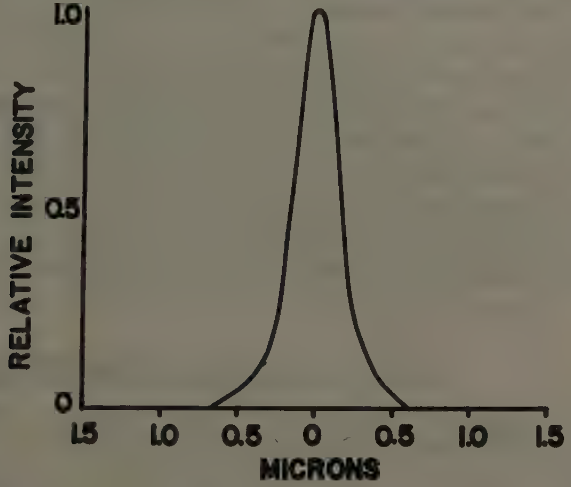


Fig. 1. Line spread function of CYDAC. The function was calculated from the analogue response of light intensity as the flying spot crossed a sharp edge on a microscope slide coated with aluminum (100x Plan Objective, NA = 1.25; Condenser NA = 0.8; wavelength centered on 500 nm). The modulation transfer function corresponding to this line spread function has been given elsewhere (5).

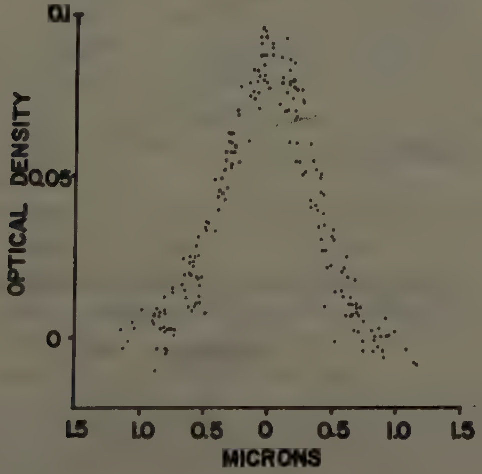


Fig. 2. Composite of 27 chromatid profiles. The profiles come from chromatids lying vertically or horizontally to the scanning raster in the digital image of a metaphase. Each profile contains approximately 10 optical densities taken at 0.25 micron intervals. The profiles have been centered on the zero line of the horizontal scale.

An Operational Definition of the Boundary

It is clear from Figure 2 that many grayness levels could serve as thresholds to convert the chromatid into a binary image. However, the width of this image would then be arbitrary, as would the average optical density of the interior. What is more, the same threshold applied to a chromosome that is more densely organized than the ones shown would intersect the bell differently and would cast doubt on the usefulness of interchromosomal comparisons of area and concentration. To illustrate this further, the effect of threshold on the area of representative chromosomes from one metaphase is shown in Figure 3. The differences in slope are significant and probably represent differences in sharpness or compaction of the individual chromosomes. In the absence of a definitive anatomical boundary, a method must be devised to set thresholds as objectively as possible. We have chosen to define the inflection points of the bell as the operational boundary (9, 13).

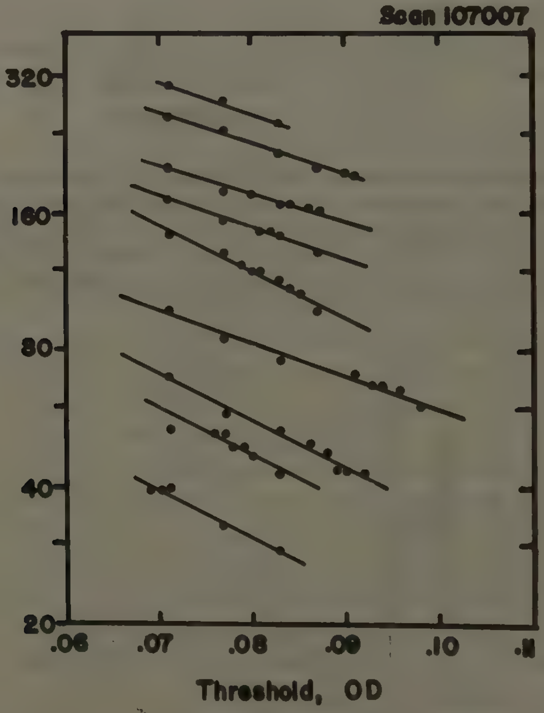


Fig. 3. Core Area as a function of threshold. Nine chromosomes, representing the full range of chromosome size, were taken from the same scan. Core Area is the area of the binary image generated by a particular threshold. The response of Core Area to the limited range of changes of threshold shown here is roughly linear on this semi-logarithmic plot. The divergent slope of several of the chromosomes is due to either focus or compaction effects. For the average-sized chromosomes, Core Area changes by almost 3 samples (0.19 square micron) for a change of one CYDAC unit (0.001 OD) in threshold.

In principle one could establish the inflection points line by line or point by point throughout the image. In the latter case this is equivalent to finding the largest local gradients of grayness on the image surface in the region of a provisional boundary. Because of the local effects of noise and the difficulties in finding and connecting local directional maxima, we turned instead to a chromosome-wide approach to the problem.

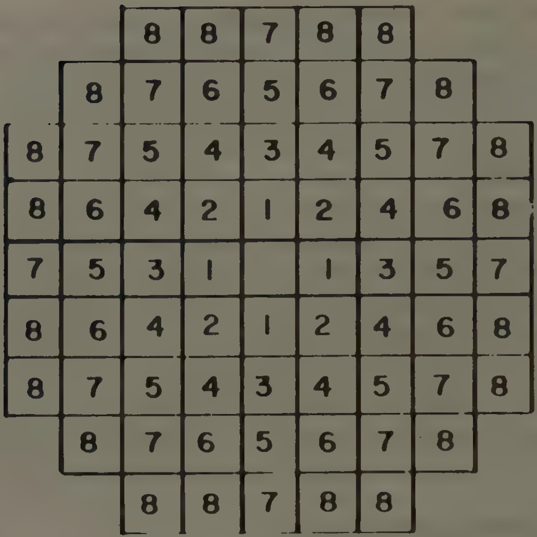
In this approach we seek a single threshold that will be appropriate for the entire chromosome using the following algorithm. The grayness information throughout the chromosome is organized in terms of distance from the edge produced by a provisional threshold, all optical densities in the same distance range are averaged, and then gradients are calculated from these mean values. If the maximum gradient is displaced from the provisional boundary, an improved threshold is calculated from the magnitude of the displacement and the steepness of the gradient. The process is iterated until it stabilizes, at which point the definitive boundary coincides with the location of the maximum gradient.

Chromosome Shells and Interaction Points

To begin the process of setting individual thresholds, a provisional overall threshold is applied to the digital image of a metaphase. This threshold partitions the metaphase into a binary image of blobs or Cores which, if a reasonable threshold has been chosen, will correspond to the individual chromosomes. Minor aberrations such as fragmented or coalescent chromosome Cores are corrected by editing at this stage (6, 7). The image is then further partitioned into successive Shells (or layers) inside and outside of each Core. All points one unit outside of Core are in Shell -1, all points two units outside are in Shell -2, and so on through Shell -8. Points beyond eight units are labeled Background. The inside points are assigned from Shell $+1$ to Shell $+8$ in the same manner (i.e., all points one unit inside Core are in Shell $+1$, etc.). Points inside of Core but greater than eight units from the Core edge are considered to be in Shell $+8$. Distance is defined by the matrix in Figure 4, and a distance unit is one-half of the sampling interval of the digital image (or one-half of 0.25 micron under our standard conditions of scanning chromosomes).

The shelling procedure parcels out the space between adjacent Cores by assigning each shelled point to the closer (or the closest) Core. These assignments and the Cores themselves are identified with an object number which either has been designated previously using punched cards or is generated by the program when the Cores are first created. Up to this point the process is equivalent to a distance transform relative to the provisional boundary (1, 12).

When two chromatids are in close proximity, the optical density surface between them is the sum of the profiles for the individual chromatids. The shape of this surface will be a function of the separation of the chromatids, and the inflection points may no longer be an indication of the correct threshold. Thus it is essential for the success of the method to isolate those non-Core points which are under the influence of more than one chromatid Core. Such points are called Interaction Points and they are identified by the following algorithm. Beginning at each



		8	8	7	8	8		
	8	7	6	5	6	7	8	
8	7	5	4	3	4	5	7	8
8	6	4	2	1	2	4	6	8
7	5	3	1		1	3	5	7
8	6	4	2	1	2	4	6	8
8	7	5	4	3	4	5	7	8
	8	7	6	5	6	7	8	
		8	8	7	8	8		

Fig. 4. The distance matrix. The numbers are the distance between the box they lie in and the nearest boundary of the center box of the grid. For chromosome scans the boxes are 0.25 micron apart and one distance unit should be 0.125 micron.

Shell -6 and Shell -7 point, the shortest path to Core is traced and each point on the path is marked as in Figure 5. When this is completed, all points in Shells -1 to -5 that are unmarked are labeled Interaction Points. According to this definition, Interaction Points either lie between Cores that are within 12 distance units (1.5 microns) of each other or lie within concave regions (such as the notch between chromatids) of a single Core.

The procedure up to this point is summarized in Figure 6.

Shell Density and Shell Gradients

Shell OD (i, j) is defined as the mean optical density of all non-Interacting Points with the i^{th} Shell assignment from the j^{th} object. Shell OD $(0, j)$ is the provisional or prevailing threshold value for the j^{th} chromosome. The Shell Gradients are the successive differences in Shell OD of adjacent Shells:

$$\text{S-Grad}(i - 0.5, j) = \text{Shell OD}(i, j) - \text{Shell OD}(i - 1, j).$$

Because of the restriction imposed by a finite rectangular grid, the units of distance in the matrix in Figure 3 contain systematic biases. These are particularly troublesome near the center of the matrix and they lead to unequal spacing of the Shells and spatial anisotrophy. Since the Shell Gradients are sensitive to these effects, they must be corrected by estimates of the true distance between the centers of successive Shells. The thickness of a particular Shell should be reflected in its area or equivalently by the number of points within it. That this is not a well behaved relationship with the rectangular grid is indicated by the Shell Areas

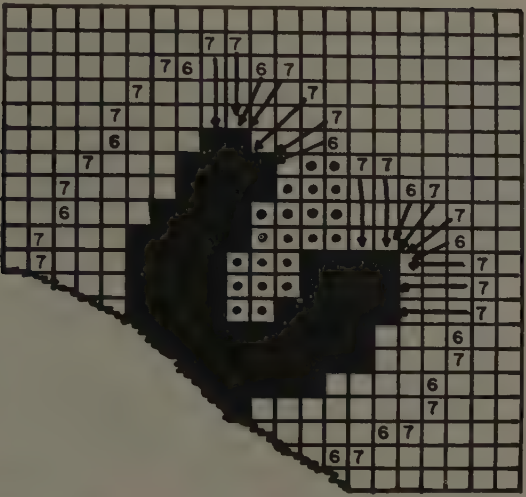


Fig. 5. Interaction Points. The blacked-in region indicates a portion of the Core of a chromosome with partially separated chromatids. The Shell −6 and −7 points are indicated by the numbers and the arrows (shown only near the tips of the chromatids) show the shortest path from these points to Core. Interaction Points are marked with a dot and are those Shell −1 to −5 points untouched by the arrows. By avoiding the Interaction Points, the generalized grayness profile can be cumulated exclusively for those parts of the object that are influenced by only one region of Core.

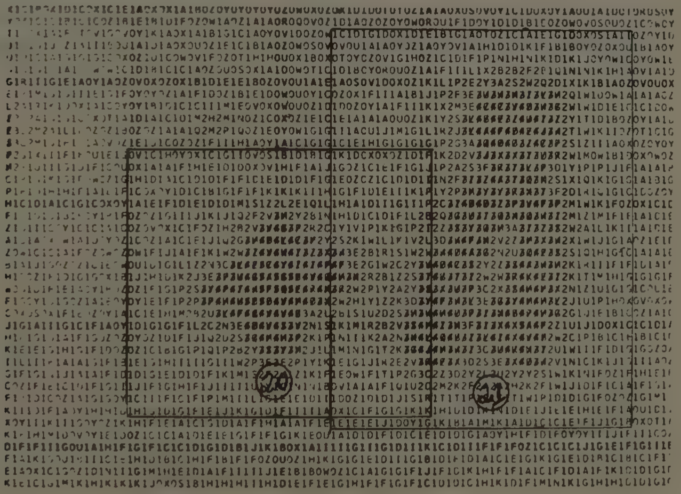


Fig. 6a. The provisional Cores and CYDAC values for two chromosomes. In this format a number and adjacent letter represent a CYDAC value. The letters A through Z omitting O give a scale of zero to 24, and the preceding number adds the corresponding multiples of 25. Thus OB is 1, OC is 2, 1D is 28, and 2E is 54. The slashes indicate that the points are above threshold and the boxes define the regions from which Background points are sampled for each chromosome.

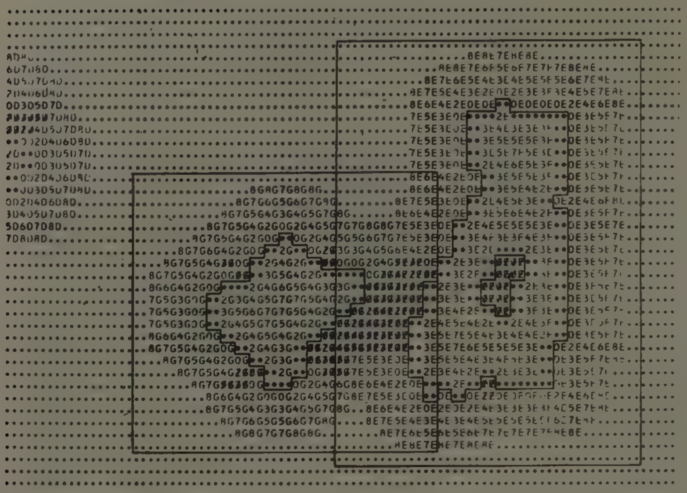


Fig. 6b. The Shell designations and Interaction Points for the same two chromosomes. The shelled points for each chromosome are labeled respectively G and E preceded by a number. For Shells plus or minus two through eight this number is the actual Shell number. Shell +1 is shown as a double asterisk and in Shell −1 the number is 0. Interaction Points are indicated by the slashes.

shown in Figure 7. These data come from 26 chromosomes divided equally between two metaphases of differing compaction. The chromosomes range widely in size and orientation and they were selected only by the criterion that they did

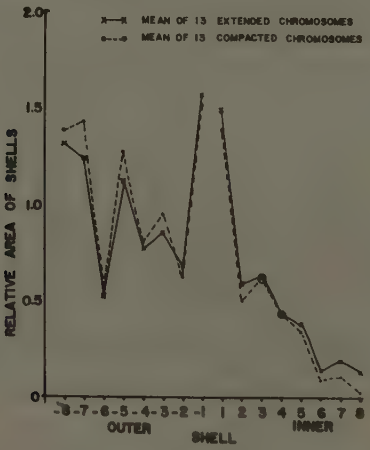


Fig. 7. Relative Shell Areas. The data are from two metaphases and five scans. Each point is the mean of the relative areas of a particular Shell from thirteen chromosomes. The actual Shell Areas or numbers of points have been normalized by the total area of each chromosome. Interaction Points are included. The highly structured appearance of the graph is due to both the inherent distortions of the distance matrix and the size, shape, and orientation of the chromosomes.

not interact with other chromosomes. The figure shows two effects: the local peculiarities due to the grid, and a general slope which reflects the expected decrease in area as the Shells shrink inward toward a point. The data fit well to a model of the chromosome Core in which the ends are semi-circles and the centers are rectangles of varying length. Using the model one can calculate the increment in thickness that corresponds to each increment in area. The distance between Shells is based on these increments in thickness and is calculated to the midpoint of each Shell. Table I shows the distance between successive centers [$D(i - 0.5)$] and the location of the center for the various Shells.

A Corrected Shell Gradient can now be calculated:

$$\text{Corr. S-Grad}\,(i - 0.5, j) = \frac{\text{S-Grad}\,(i - 0.5, j)}{\text{D}(i - 0.5)}$$

The Corrected Gradients are smoothed:

$$\text{Sm.Corr. S-Grad}\,(i - 0.5\,j) = 0.25\,[\text{Corr. S-Grad}\,(i - 1.5, j) + 2 \cdot \text{Corr. S-Grad}\,(i - 0.5, j) + \text{Corr. S-Grad}\,(i + 0.5, j)].$$

Table 1. Average Shell Dimensions for Chromosomes

Shell	*Effective Distance from Center of Shell to Boundary **	*Effective Distance Between Successive Centers **
−8	7.669	
		1.166
−7	6.503	
		0.840
−6	5.663	
		0.822
−5	4.841	
		0.968
−4	3.873	
		0.866
−3	3.007	
		0.830
−2	2.177	
		1.268
−1	0.909	
		0.909
0	0.000	
		0.964
1	0.964	
		1.375
2	2.339	
		0.919
3	3.257	
		0.897
4	4.155	
		0.743
5	4.897	
		0.470 **
6	5.368 **	
		0.286 **
7	5.654 **	
		0.262 **
8	5.916 **	

* Distances are in units defined by the matrix in Figure 1. One such unit equals 0.125 micron at the magnification of 800x used in CYDAC scans of chromosomes.

** These values depart from expectation because of a breakdown in the chromosome model as the Shells shrink inward. They represent probability of occurrence as well as literal distance, and they are not used in the definitive analyses.

The three Smoothed Corrected Shell Gradients on each side of the provisional boundary are fitted by the least-squares method to a quadratic curve,

$$\text{Fitted Grad} = a + bx + cx^2,$$

where x is distance from the boundary (i.e., plus or minus 0.5, 1.5, and 2.5 distance units for the six estimated gradients). The location of the Maximum Shell Gradient is at $-b/2c$, and the recommended change in threshold is $f(ab/2c - b^3/12c^2)$ where the constant f is used to compensate for the effect of smoothing and is currently set at 1.4. The magnitude of the Maximum Shell Gradient is found by setting x equal to $-b/2c$.

After each chromosome is put through the shelling procedure, all of the chromosomes are rethresholded according to the recommended values. The shelling-rethresholding cycle is then repeated as many times as desired.

Results

To illustrate a typical result, the data from a human number 6 chromosome are given in Table II and Figure 8. The Table shows the steps of the calculation with the data expressed in CYDAC and distance units (where 100 CYDAC units equals 0.1 OD, and 1 distance unit equals 0.125 micron). The Figure displays Shell OD and Corrected S-Gradients in conventional units as a function of distance from the boundary. The Shell OD curve is the generalized grayness profile for this chromosome. It shows the asymmetrical sigmoidal quality that characterizes the profile of every chromosome we have measured. Shells 3 and 4 correspond to the axis of the chromatids. Beyond this the generalized grayness profile of the interior Shells is not predictable because points this far into the chromosome are divided

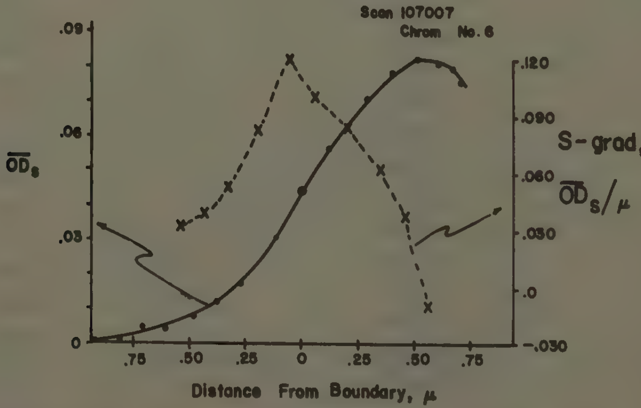


Fig. 8. Generalized grayness profile and Smoothed Corrected Shell Gradients of a chromosome. The data are given in OD and micron units, and OD_s refers to Shell OD. Calculations related to the figure are given in Table II. The inflection point of the generalized grayness profile lies very close to the provisional boundary in this example, and the asymmetrical sigmoidal shape of the curve is characteristic of chromosome profiles in general. The sharp peaking of the Shell Gradients is also typical.

Table 2. Intermediate and Derived Data from One Chromosome

Shell	Shell Area	Shell OD	Shell Gradient	Corrected Shell Gradient	Smoothed Corrected Shell Gradient	Fitted Shell Gradient
−8	37	0.99				
			0.34			
−7	45	1.33				
			3.44	4.26		
−6	20	4.77				
			−0.52	−0.66	1.71	
−5	40	4.25				
			3.79	3.90	2.99	
−4	26	8.04				
			4.19	4.82	5.04	
−3	31	12.23				
			5.30	6.62	7.10	7.11
−2	19	17.53				
			13.06	10.32	10.55	10.80
−1	48	30.59				
			13.57	14.93	13.23	12.67
0		44.16				
			12.23	12.72	12.72	12.71
1	67	56.39				
			14.38	10.50	10.35	10.94
2	23	70.77				
			7.06	7.69	7.64	7.34
3	37	77.83				
			4.22	4.68	3.94	
4	16	82.05				
			−0.96	−1.29	−0.49	
5	18	81.09				
			−1.87	−4.05	−5.56	
6	5	79.22				
			−3.68	−12.83		
7	8	75.55				
			−7.13			
8	1	68.42				

Location of Maximum Shell Gradient	0.025 distance unit
Recommended Change in Threshold	0.46 CYDAC unit
Maximum Shell Gradient	12.92 CYDAC unit/distance unit

between the high densities of closely interacting chromatids and the low densities or troughs between those chromatids which are at some distance from each other. Although the highest Corrected Shell Gradient for this particular chromosome is outside of the boundary, after smoothing and quadratic fitting the Maximum Shell Gradient lies 0.025 distance units inside and hence the recommended threshold is 0.46 CYDAC units higher than the one used.

To test the convergence of the thresholding procedure, a digital image of a metaphase was thresholded at three arbitrary levels and each of these was then used as the provisional start for a sequence of analyses. A representative sample of the resulting chromosome thresholds is summarized in Figure 9. Initially the arbitrary thresholds for the 26 chromosomes were 6 CYDAC units apart. By the end of the first cycle this difference was reduced to a mean of 0.13 CYDAC units and it then fell to 0.05 CYDAC units at the end of the second cycle. The corresponding standard deviations of the differences for individual chromosomes were 3.17 and 1.39 CYDAC units. Perhaps the most striking aspect of Figure 9 is the different levels at which each chromosome threshold stabilized.

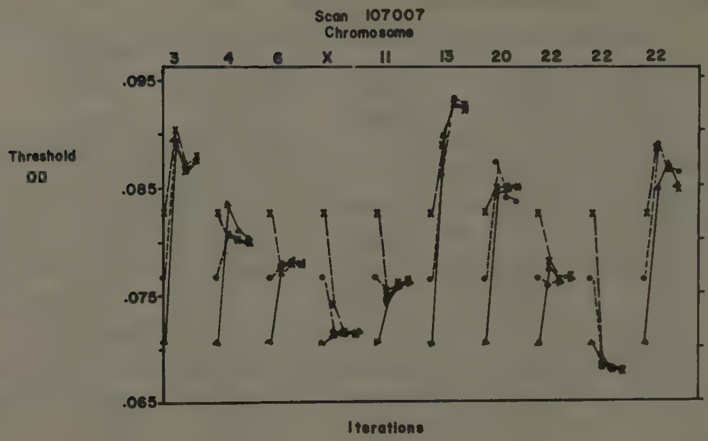


Fig. 9. Convergence of the thresholds for a single image. Ten representative chromosomes are shown from a single scan of a metaphase. The chromosome numbers are given at the top of the figure. At the outset three overall thresholds were used at 6 CYDAC unit (0.006 OD) intervals. After one pass through the shelling-rethresholding procedure the effect of the starting threshold had been largely eliminated, and with subsequent passes the threshold for all three experiments stabilized to the same value. Note the striking difference in threshold for different chromosomes from the same cell.

It is clear from the preceding test that convergence to within 1 or 2 CYDAC units (0.001 or 0.002 OD) can be expected for the digital image of a single chromosome. But it is not yet clear how stable the resulting threshold is from one image to the next, how this affects the measurement of parameters such as area, and how meaningful the grossly different thresholds are for chromosomes lying side by side. A more stringent test of the method is the comparison of results when the same material is scanned more than once.

In Table III the means and coefficients of variation are given for four parameters from paired scans in which the positions of the 21 chromosomes were unchanged between scans and great care was exercised to maintain optimal focus throughout. The replication error of 2.6 percent for Integrated Optical Density is similar to values we have published previously (7) as is to be expected since Integrated Optical Density is insensitive to changes in the threshold level. The threshold itself, shown here as Shell OD (0, *j*), has a replication error which is only slightly larger than that found in the preceding convergence experiment. Again the results indicate a highly significant difference in threshold among chromosomes within the same cell. Core Area has a replication error of 5.2 percent which is roughly twice the repeatability of Integrated Optical Density, and is the square of the replication error of threshold. The Maximum Shell Gradient is a measure of the sharpness of the generalized grayness profile. It also shows a small replication error and highly significant differences among chromosomes within the same cell.

Table 3. Measurements on 21 Chromosomes after Three Passes Through the Shelling-Rethresholding Procedure

Parameter	*Units*	*Mean*	*Replication Error **	*Interchromosomal Variability ***
Integrated OD	$(\mathrm{OD})\mu^2$	0.466	2.6%	42%
Core Area	μ^2	7.53	5.2%	48%
Shell OD (0,j)	OD	0.0303	2.3%	13.5%
Maximum Shell Gradient	OD/μ	0.0856	4.3%	11.2%

* Coefficient of variation of replication of 21 pairs.
** Coefficient of variation of 21 measurements.

Figure 10 is a summary of the generalized grayness profiles of 24 human chromosomes. These come from two metaphases involving five scans, and they cover the full range of chromosome size and compaction. They have been normalized for each chromosome by calculating the Shell OD as a percentage of the Mean of Shell OD 3 and 4. As is indicated by the narrow 95 percent confidence limits throughout the curve, there is a remarkable degree of conformity both within and between metaphases.

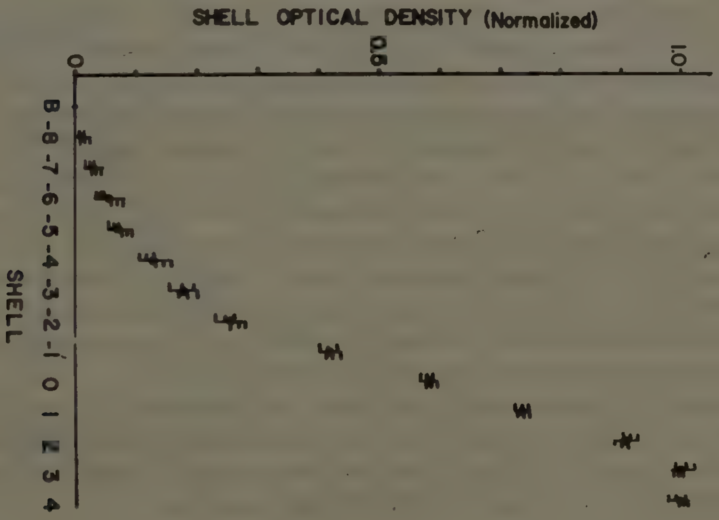


Fig. 10. The generalized grayness profile for 24 chromosomes from two human metaphases. For each chromosome the Shell OD was calculated as a percentage of the mean of Shell OD 3 and 4. The percentages were then combined into means and 95 percent confidence limits for each of the two metaphases involved. The horizontal bars on the left of the symbols refer to the mean and limits for the metaphase with extended chromosomes, and the bars to the right refer to the mean and limits for the metaphase with compacted chromosomes. The overall mean is shown as a dot in the center of the symbol.

Discussion

The transformation of an image into a generalized grayness profile preserves optical density while reducing the spatial dimensions from two to one. This reduction is justifiable only in an image in which the grayness response at the edge is expected to be uniform. In such an image the profile provides a convenient and effective summary of edge behavior, including the location of the inflection point and the sharpness of the edge.

A priori one would think chromosomes would be well suited for this type of analysis because of the underlying linear architecture of the chromatids. However, it was not clear at the outset to what extent the edges would be uniform and with what reliability the generalized grayness profile could be extracted from the several hundred optical density values constituting a digital chromosome image.

Our initial results with chromosome images indicate that the generalized grayness profile and the parameters derived from it are highly reproducible and well behaved. The uniformity of the profiles from a variety of chromosomes suggests that there is a surprising degree of homogeneity in the apparently haphazard cross-sectional architecture of chromatids. The generalized grayness profile may provide some insight into the actual structure of chromosomes, and we have already found it to be a convenient and powerful tool for the enhancement and processing of chromosome images.

We anticipate that the stability of parameters derived from the generalized grayness profile will make a significant contribution to the resolution of differences among chromosomes within metaphases. Experiments currently in progress indicate that these parameters can be extremely sensitive to focus and compaction. In relation to focus this has already suggested methods to measure optimal focus at the time of scanning, and in relation to compaction there is the possibility of using the generalized grayness profile to estimate compaction directly. Compaction may have some discriminatory utility in its own right, and its measurement could improve other parameters. For example, Integrated Optical Density is insensitive to focus and to thresholding, but it is slightly and predictably affected by compaction (7). An independent measurement of compaction would allow a correction of the Integrated Optical Density and should improve discriminatory ability.

The generalized grayness profile has not as yet been tested on other kinds of images. Although we are planning several such applications, it is our hope that the ultimate proving ground for the method will be in other laboratories and on a much broader spectrum of images than we can study ourselves.

REFERENCES

1. Blum, H., A transformation for extracting new descriptors of shape. (In) *Models for the Perception of Speech and Visual Form,* M.I.T. Press, 1967.
2. Caspersson, T. *Cell Growth and Cell Function,* Norton, New York, 1950.
3. DuPraw, E. J. and G. F. Bahr., The arrangement of DNA in human chromosomes, as investigated by quantitative electron microscopy. *Acta Cytologica* **13,** 188-205, 1969.
4. Mayall, B. H. and M. L. Mendelsohn. Errors in absorption cytophotometry: Some theoretical and practical considerations. (In) *Introduction to Quantitative Cytochemistry,* Volume 2 (G. L. Wied and G. F. Bahr, eds.) Academic Press, 1970. In Press.
5. Mendelsohn, M. L., B. H. Mayall, J. M. S. Prewitt, R. C. Bostrom, and W. G. Holcomb., Digital transformation and computer analysis of microscopic images. (In) *Advances in Optical and Electron Microscopy,* Volume 2 (R. Barer and V. E. Cosslett, eds.), Academic Press, 1968, pages 77-150.
6. Mendelsohn, M. L., B. H. Mayall, and J. M. S. Prewitt. Approaches to the automation of chromosome analysis. (In) *Image Processing in Biological Science* (D. M. Ramsey, ed.), UCLA Forum in Medical Science No. 9, 1968, pages 119-136.
7. Mendelsohn, M. L., D. A. Hungerford, B. H. Mayall, B. H. Perry, T. J. Conway, and J. M. S. Prewitt. Computer-oriented analysis of human chromosomes II. Integrated optical density as a single parameter for karyotype analysis. *Ann. N. Y. Acad. Sci.* **157,** 376-392, 1969.
8. Mertz, P. and F. Gray. A theory of scanning and its relationship to the characteristics of the transmitted signal in telephotography and television. *Bell System Tech. J.* **13,** 464-515, 1934.
9. Osterberg, H. and L. W. Smith. Optical measurement of self-radiant line elements by the method of steepest slopes. *J. Opt. Soc. Am.* **54,** 406-410, 1964.
10. Prewitt, J. M. S. and M. L. Mendelsohn., The analysis of cell images. *Ann. N. Y. Acad. Sci.* **128,** 1035-1053, 1966.
11. Roberts, F. and J. Z. Young. Scanning microscopy. *Proc. I. E.E.* **99,** Part IIIA: 747, 1952.
12. Rosenfeld, A. and J. L. Pfaltz., Sequential operations in digital picture processing. *J.A.C.M.* **13,** 471-494, 1966.
13. Welford, W. T., The Mach effect and the microscope. (In) *Advances in Optical and Electron Microscopy,* Volume 2 (R. Barer and V. E. Cosslett, eds.), Academic Press, 1968. pages 41-76.

A MODEL OF A COMPUTER-BASED SYSTEM FOR MEDICAL RECORD-KEEPING BY PHYSICIANS

Robert A. Greenes*

Introduction

The mutual effects of technology and society are receiving consideration in multiple spheres (1,2); these issues are perhaps nowhere of more concern than in the present structure of medical care. The continuing series of advances in scientific and technological prowess, and in social attitudes toward availability of medical care, have, despite their positive effects, contributed to major deficiencies in the mechanisms for delivery of medical care.

Paradoxically, there is both an increased amount of information to be used in the treatment of disease, and an increased discontinuity in the sources of this information. The difficulties and the costs of communicating and coordinating this information have multiplied accordingly (3). Most clinical information is presently handwritten, often in illegible script on multi-part forms. Most of the information can be retrieved only by shuffling through stacks of reports or by leafing through the medical record. A recent study of hospital costs has shown that over 25% of the hospital budget is spent on information processing (4). Physicians, nurses, and technical staff—medicine's scarcest resource—spend a significant proportion of their time in bookkeeping activities.

The diagnostic and therapeutic decision-making process ideally occurs in a situation in which the physician's knowledge base intersects significantly with the information base that characterizes his patient at a particular moment in time. There are two major difficulties that mitigate against this ideal situation. First, the knowledge base of physicians must of necessity become increasingly narrow in scope as the scientific and technological mastery of a specialty area becomes

---

* Supported in part by grant HM-688-01 from the National Center for Health Services Research and Development and in part by contract PH-43-68-1343 with the National Institutes of Health.

increasingly difficult. This has indeed occurred, as a result of which the patient must seek medical care from a variety of physicians, each of whose knowledge base intersects best with his current problem. As a consequence, however, the information base about the patient becomes fragmented, and it becomes increasingly difficult for the individual physician to appreciate the total status of his patient.

Secondly, the physician is dependent upon many supporting services and ancillary events in order to obtain the information he requires. Although he is responsible for setting most of these activities in motion, the information from each of these is derived independently from him, and must be made available to him for his decision-making process.

Thus the need for communication and coordination of activities involved in the process of medical care is vast. Many of the inefficiencies and much of the expense of medical care may be attributed to lack of communication, delays, or duplication of effort, or the inability to make proper decisions because of the lack of necessary data. The discontinuities engendered by the advances of the past century have created an urgent need for development of appropriate systems for the management of clinical data. We may identify a number of objectives which an ideal clinical data management system must facilitate:

(1) review of clinical information for patient care;
(2) quality control and auditing of the medical care process;
(3) clinical investigation and record research;
(4) administrative record management; and
(5) administrative planning, allocation of resources, and utilization review.

In order to meet these objectives, the information that comprises a clinical data management system must possess a number of attributes, among which are: (1) legibility, (2) consistency of terminology, (3) organization and accessibility, (4) codability for retrieval, (5) compactness and conciseness of storage.

The only clinical data management system that currently exists is the conventional paper-bound patient record, although the objectives above are fulfilled only to a minimal extent by it. Despite a changing role and increased demands on it, the medical record has evolved little in form over the past century. Medical records remain a chronological repository of data; they possess no organization by diagnostic or therapeutic problem. Relevant records may frequently not be located. Notes pertinent to a particular aspect of a patient's health may be accessed only by leafing through the entire volume. Terminology is inconsistent, and notes are often illegible.

Computer-based medical record systems have been proposed as a solution to many of the problems of the conventional medical record. Of particular concern in this discussion is the narrative information recorded by the physician, in the form of notes on the patient's history and physical examination, progress notes, consultations, etc. Most of the decision-making activities of medical care are documented in narrative fashion, and much of the data on which it is based is, or should be, indicated in the physician's note. Nevertheless, this "soft" data has been in the past least amenable to the introduction of automated techniques which would achieve the objectives that have been itemized above. Two other sources of data have yielded quite successfully to the introduction of automated techniques for data acquisition: (1) the patient himself, and (2) the clinical laboratory. Automated patient interviewing techniques have been developed and tested extensively, and the evidence for patient cooperation, physician acceptance, and accuracy and reliability of the information collected have been convincingly demonstrated (5, 6, 7). Clinical laboratory information processing has been the area in which the application of computers has been most successful to date. Numerous systems are either under development or operating in hospitals and other institutions throughout the United States (8, 9, 10, 11), and their use will be routine for hospitals of almost any size within a few years. The need for such systems has long been recognized, and because the information handling tasks have been relatively straightforward, this development is quite far advanced currently.

While systems for patient history taking and clinical laboratory processing constitute valuable and essential inputs for a medical record system, the primary focus of a clinical data management system must be the physician. A number of investigators have experimented with alternative ways of capturing information from the physician. These have ranged from the entry of free text narrative information directly into the computer in its original form (12, 13), to the use of highly structured, detailed check lists for recording observations by physicians in a compact, coded format (14, 15, 16). In general, the bulkiness, redundancy, and lack of retrievability of narrative text information have made the first method unsatisfactory. The applicability of the second approach has been limited by the contradiction between the size of a check list required if it is to be sufficiently detailed to allow a physician to record his findings in the depth that he desires, and the brevity required if physicians are to be willing to use it.

A compromise between the flexibility, for recording of free narrative text input and the use of structured check lists lies in the development of systems which support on-line conversational interaction between the physician and the computer. Such systems permit a high degree of structure of content, while providing the user with the option of selecting only those topics which he wishes to pursue. An additional advantage lies in the feedback that can be built into an interactive system as a means of providing quality control and didactic material. It is important to recognize that much of the information that is recorded in the medical record is not really unconstrained free text, but rather patterned statements that the physician usually uses to describe particular ideas (17). Considering the diversity of individuals in medicine and of their backgrounds and training, there is remarkable consistency in the way in which particular ideas tend to be expressed. This is perhaps best illustrated by the recording of the results of a physical examination, a process which usually consists of the selection of a small number of fairly stylized phrases from a large repertoire of possible descriptions of findings. A branching dialogue with the computer may facilitate this process by permitting the physician to select very quickly the topics which he wishes to discuss. These topics may be expanded in further displays, and material that is not relevant to the current problem may be bypassed. The use of structured displays and directed conversation not only provide standardization of terminology, but also may improve the comprehensiveness with which information is recorded, in that a display may remind a physician to review topics that he may otherwise have neglected to record.

Kiely et al. (18) and Slack et al. (19) have reported the development of interactive computer programs for the acquisition of clinical information from physicians by means of a branching question/answer dialogue with a television-like, cathode ray tube (CRT) display terminal. Neither of these programs has been sufficiently complete to be evaluated as part of an operational medical record system, although initial experiences have indicated the potential usefulness of the approach. Further impetus has been provided by the work of Weed (20), who has advocated a problem-oriented approach to recording information in the medical record; a significant aspect of this approach lies in its application to automated clinical data acquisition. An interactive system might, for example, rely on the use of a hierarchical structure of information, in which subsequent displays are presented to the physician only if he selects for discussion a particular branch or category in the hierarchy. The problem-oriented approach allows the physician to use such a branching interactive dialogue with the computer in a logical and natural manner.

Although much of the initial experimental effort in the exploration of the interactive approach has been concerned with the development of the technology utilized for the input of clinical information, it is necessary to view this approach in relation to a full-scale system of which it might form the basis. A model of a computer-based medical record system based on physician/computer interaction needs to be developed which will provide (1) detailed specifications of the methods by which the various objectives of a clinical data management system can be fulfilled, and (2) a rational basis for determining the manner in which the various aspects of the model may be implemented. Weed (20) has emphasized the necessity for building a clinical data base organized to represent the current state of medical knowledge, as the basis for a medical record system. While this is an extremely valuable aspect of the development, it is a long range effort that must be based soundly upon an understanding of the way in which that data base will form part of an operational medical record system. While research workers in this field may feel strongly that interactive data acquisition is the only way that medical record keeping can reasonably hope to meet the objectives we have stated, carefully designed experiments and evaluation efforts must be undertaken to demonstrate its feasibility. The intermediate tangible results of a functioning system in a limited area would be very valuable to a continuing effort of this type.

Accordingly, this paper will describe a model for clinical data management, and indicate the issues that must be considered in its implementation. These considerations are based on experience with a prototype implementation of the model, which has been evaluated under routine use in a limited clinical setting. Where pertinent, results of this evaluation effort will be indicated in the subsequent discussion; however, this prototype implementation is described in detail elsewhere (21, 22).

A Model for a Clinical Data Management System

We will consider the model from the point of view of continuing medical care, since it is this aspect of clinical medicine that is most concerned with the data in the medical record. The information from initial patient workups can be introduced into a computer system in a variety of ways, and serve to establish a baseline data base for the patient in a computer-based patient file. All subsequent medical care is documented as changes to this data base, always in the form of additional findings or changes in status of old findings.

Fig. 1 illustrates the operation of a medical record system for facilitating continuing care in a management oriented specialty clinic. The prototype system that has actually been implemented follows this design quite closely. We have chosen to focus initially on the Hypertension Clinic in the Massachusetts General Hospital Outpatient Department. Reasons for selection of this clinic were twofold: First, the

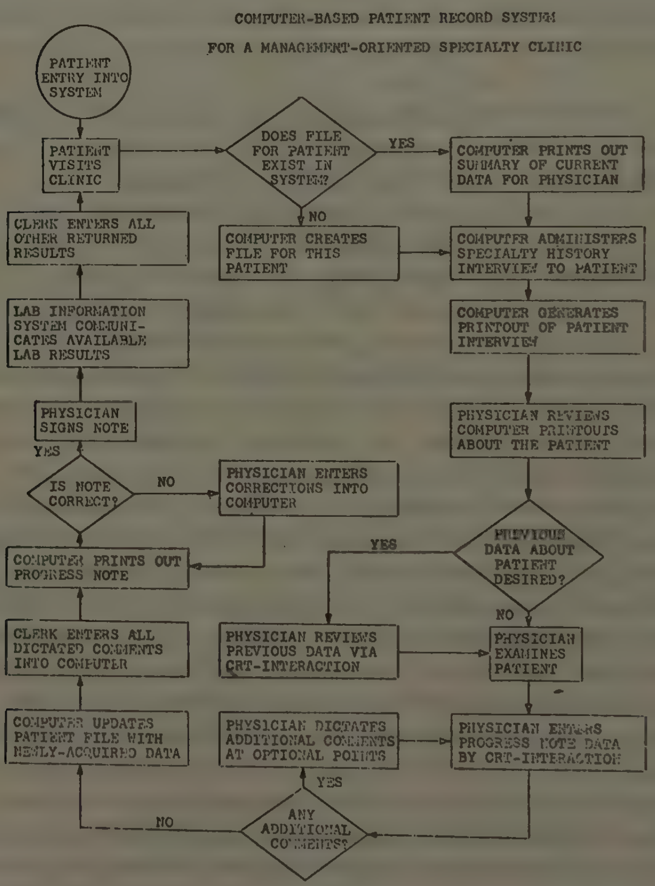

COMPUTER-BASED PATIENT RECORD SYSTEM
FOR A MANAGEMENT-ORIENTED SPECIALTY CLINIC
PATIENT ENTRY INTO SYSTEM
PATIENT VISITS CLINIC
DOES FILE FOR PATIENT EXIST IN SYSTEM?
YES
COMPUTER PRINTS OUT SUMMARY OF CURRENT DATA FOR PHYSICIAN
NO
CLERK ENTERS ALL OTHER RETURNED RESULTS
COMPUTER CREATES FILE FOR THIS PATIENT
COMPUTER ADMINISTERS SPECIALTY HISTORY INTERVIEW TO PATIENT
LAB INFORMATION SYSTEM COMMUNICATES AVAILABLE LAB RESULTS
COMPUTER GENERATES PRINTOUT OF PATIENT INTERVIEW
PHYSICIAN SIGNS NOTE
PHYSICIAN REVIEWS COMPUTER PRINTOUTS ABOUT THE PATIENT
YES
IS NOTE CORRECT?
NO
PHYSICIAN ENTERS CORRECTIONS INTO COMPUTER
YES
PREVIOUS DATA ABOUT PATIENT DESIRED?
COMPUTER PRINTS OUT PROGRESS NOTE
NO
CLERK ENTERS ALL DICTATED COMMENTS INTO COMPUTER
PHYSICIAN REVIEWS PREVIOUS DATA VIA CRT-INTERACTION
PHYSICIAN EXAMINES PATIENT
COMPUTER UPDATES PATIENT FILE WITH NEWLY-ACQUIRED DATA
PHYSICIAN DICTATES ADDITIONAL COMMENTS AT OPTIONAL POINTS
PHYSICIAN ENTERS PROGRESS NOTE DATA BY CRT-INTERACTION
YES
NO
ANY ADDITIONAL COMMENTS?

Fig. 1. A schematic flow chart of the operation of a computer-based medical record system in a management-oriented specialty clinic.

patients in this clinic have problems that define a sufficiently circumscribed subset of internal medicine that it was considered readily feasible to structure the medical content of this specialty area. Because of this limitation of content area, progress notes on most patients in this clinic could be recorded by selection of choices from the structure. Secondly, because this is a management oriented clinic, operation of the system in this environment afforded an excellent opportunity to underscore the utility of a computer based medical record system for providing continuity and comprehensiveness of patient care. Aspects of the model that were implemented in the prototype included: (1) total recording of clinical information in the computer system by physician/computer interaction, (2) storage of this information in a centralized computer-based file, (3) hard copy print-out of progress notes for the conventional medical record, (4) production of summaries and tabular displays to aid in review of the medical record, (5) on-line interactive retrieval and review of clinical notes on particular problems for individual patients, (6) research retrieval capabilities for clinical investigation, and (7) the ability to audit and assess the quality of care as reflected in the record. A secondary objective in the use of this system was to provide a teaching instrument, which would facilitate the training of young physicians in determining what data to record. Concomitantly, it was desired to design a system that could be used expediently by the senior staff physician as well.

Data Base of Medical Content

The essential feature of the model is a tree-structured data base of terms and phrases necessary to describe the clinical status of a patient. The hierarchical organization follows a problem orientation, so that particular problems may be described by selection and following of appropriate branches from the top level of the tree downward. As one descends the hierarchy, the content becomes increasingly specific; thus by designating the branch or branches that are pertinent to a particular clinical situation, one automatically defines a context in which all subsequent detail that is presented has been selected for relevancy to that situation. Weed, who is now at the University of Vermont, is attempting to develop such a tree structured data base for clinical medicine (23). Another related effort is that by Gordon (24), whose objective is to reduce the redundancy and ambiguity of terminology by a process of continual refinement and cross referencing of clinical terms in a disease-oriented compendium of clinical data.

A complementary tactic that may be pursued involves data base development in a modular fashion, for particular specialty orientations, by subject matter experts in the areas concerned. Extension of this latter approach to the structure of the content of medicine might be achieved by development of similar structures concurrently for a number of medical specialty orientations. Content is best structured by subject matter experts, and at the same time is best evaluated by operational usage in those subject matter areas. Furthermore, the prospect of an operational system is a significant stimulus to participation in the subject matter development itself.

As the approach is applied to other specialty orientations, there may be considerable overlap with those areas already structured. For example, the data base concerned with the problem of hypertension is subdivided into the following areas: etiologic workup and quantification of hypertension; its management and control; and chest, peripheral vascular, ocular, central nervous system, and renal problems which may occur as complications. In this data structure, illustrated in Fig. 2, all problems which the patient might present that are not related to these categories, can be associated with the branch of the structure, labeled "other problems." Each of the categories is subdivided with respect to medical problems in that category, and still further into history of present illness of each problem, relevant physical examination data, laboratory tests and procedures, therapy, diagnostic assessment, and quantification of severity and progression of the problem. Each of these subsections is further elaborated by lists of words and phrases detailing pertinent aspects of the problem.

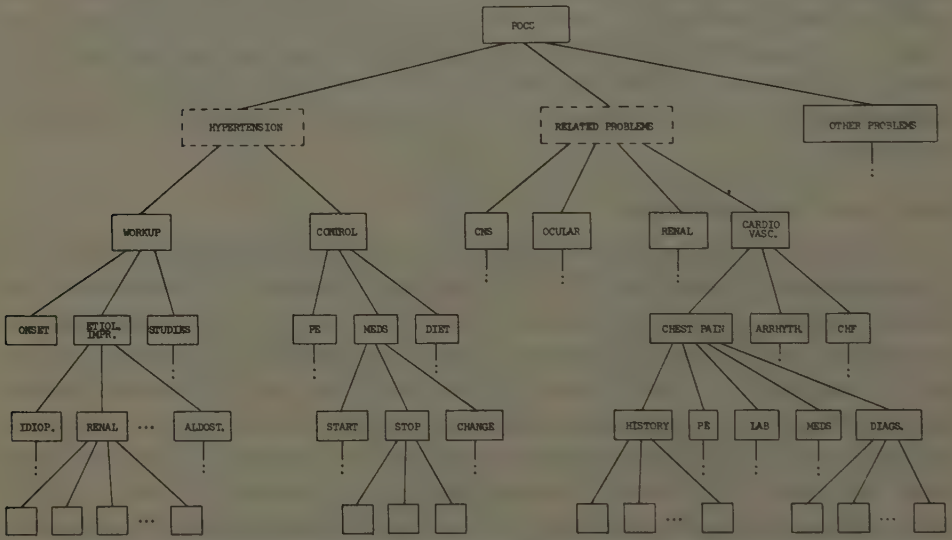


Fig. 2. Data structure used as a framework for representing the medical data describing the problems of patients in the hypertension clinic. The organization is problem-oriented hence the abbreviation POCS (problem-oriented clinical summary) at the top level node. Two main branches of the structure are concerned with hypertension workup and control. Another series of branches is concerned with the various related problems and complications of hypertension. A final branch is associated with all other problems. Further levels of detail associated with some of these branches are depicted in the figure.

The extension of this approach to a diabetes clinic might utilize much of the sections concerned with the chest, peripheral vascular, ocular, and renal problems in the hypertension data structure. While it is not easy to estimate what percentage of the subject matter of internal medicine has been structured in the data base established for the hypertension clinic experiment, subjective impressions of a number of clinicians imply that perhaps one to five percent of the content of medicine is represented. The time required to structure this perhaps 2% of medical content was approximately one man-month (three physicians part time over a three month period). Therefore, approximately fifty man months or four man years of effort might be a reasonable estimate for the amount of time required to organize the entire content of clinical medicine, even if we ignore the beneficial effect of overlap with areas already structured. Four individuals working together for a year, or more individuals for a lesser time, might accomplish this task.

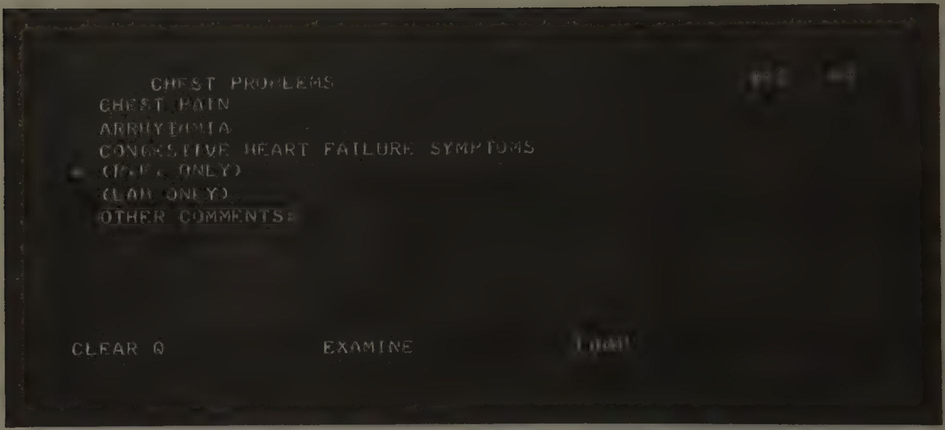
Any structure developed will reflect not only the attitudes and the orientations of its designers, but a view of medical knowledge as it exists at a particular moment in time. Therefore, while structure is essential in order to achieve the objectives we have stated, flexibility of expression must always exist. Furthermore, mechanisms must be provided for adapting the structure to changing needs so that it at all times remains relevant and appropriate. The task of maintaining a structure representative of current knowledge is fortunately a much simpler task than is its *de novo* development.

Physician Entry of Data

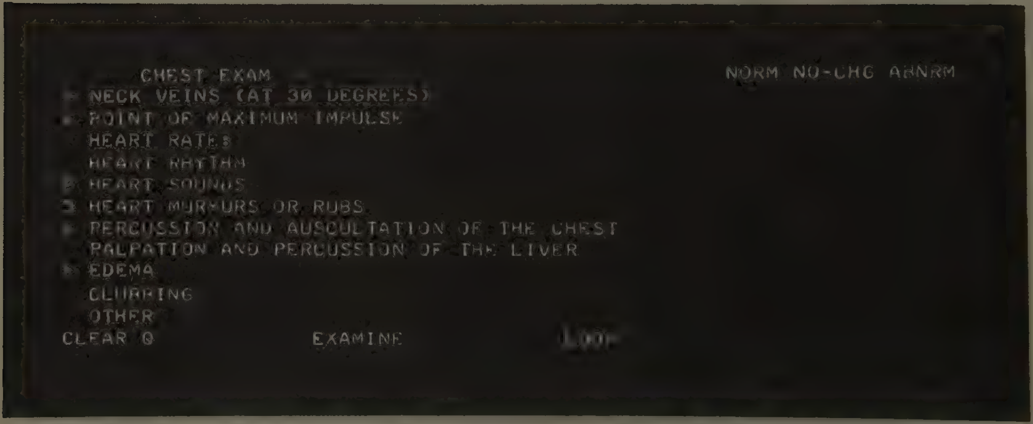
A progress note may be entered into the computer by means of an interactive session between the physician and a cathode ray tube terminal (CRT). The physician might be initially presented a list of choices, indicating the major areas or topics available to him. After the physician has made his selections, the computer processes the information received and determines whether further detail exists in the data structure for the topic selected. If this is the case, a subsequent display is presented, providing him with the list of options a level deeper in the data structure. He may thus descend the hierarchy, node by node, to an extent commensurate with the amount of detail he wishes to record. Certain branches in the data structure might take him automatically to other topics to discuss. For example, if he indicates that the patient has chest pain, he may choose to discuss history of present illness, physical examination, lab, or therapy, at his option, but if he selects the history topic, he might be presented a fixed sequence of displays regarding frequency, duration, character, location, aggravating factors, and causes of relief. Fig. 3 illustrates a sequence of displays from an actual interactive session in the Hypertension Clinic, in which the physician is recording in his progress note the results of examining the patient's chest.

Some of the substructure in the tree should be able to be reached by branching pathways from a variety of nodes, by permitting the structure to rejoin at lower levels. Once the physician has recorded all the data he desires, relevant to a particular problem, the display at the top level can be re-presented, permitting him to indicate any other areas that he may wish to discuss. In any display, the physician may make more than one choice, and if branching from more than one of these is required, the computer will follow one pathway at a time until they are all completed.

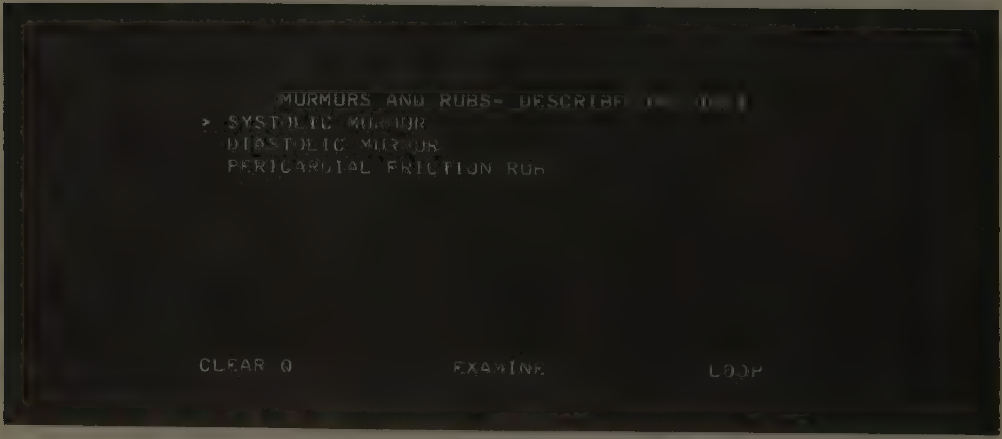
At specified points in the interaction with the data structure, opportunity should be provided for entering narrative text if desired. This might be typed by the physician on the keyboard of the display terminal, or alternatively, entered by



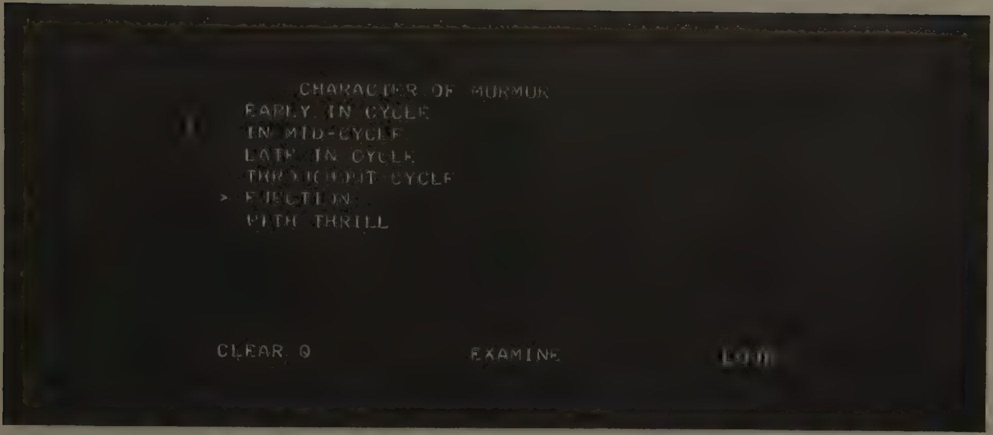
```
     CHEST PROBLEMS
   CHEST PAIN
   ARRHYTHMIA
   CONGESTIVE HEART FAILURE SYMPTOMS
 > (P.E. ONLY)
   (LAB ONLY)
   OTHER COMMENTS:

 CLEAR Q              EXAMINE              LOOP
```



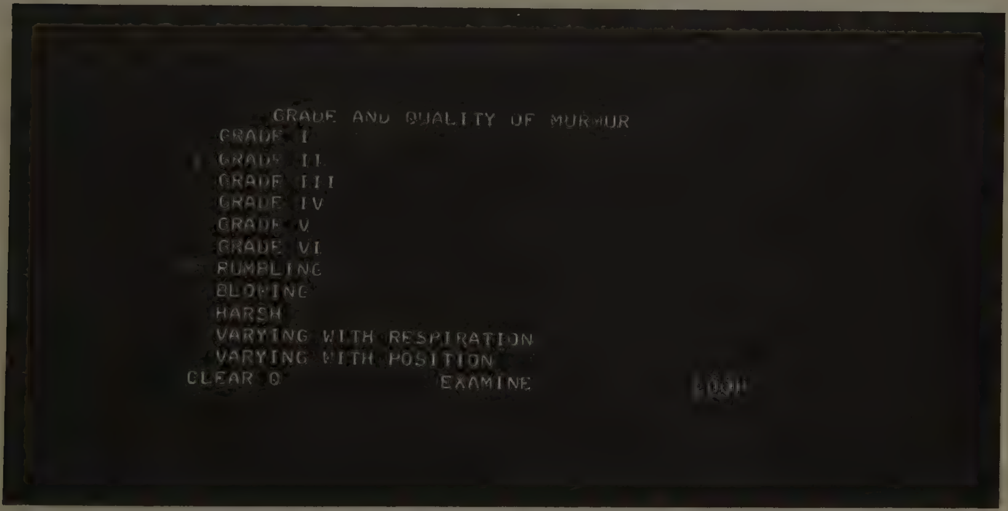
```
     CHEST EXAM                                    NORM NO-CHG ABNRM
 > NECK VEINS (AT 30 DEGREES)
 > POINT OF MAXIMUM IMPULSE
   HEART RATE
   HEART RHYTHM
 > HEART SOUNDS
 > HEART MURMURS OR RUBS
 > PERCUSSION AND AUSCULTATION OF THE CHEST
   PALPATION AND PERCUSSION OF THE LIVER
 > EDEMA
   CLUBBING
   OTHER
CLEAR Q              EXAMINE              LOOP
```



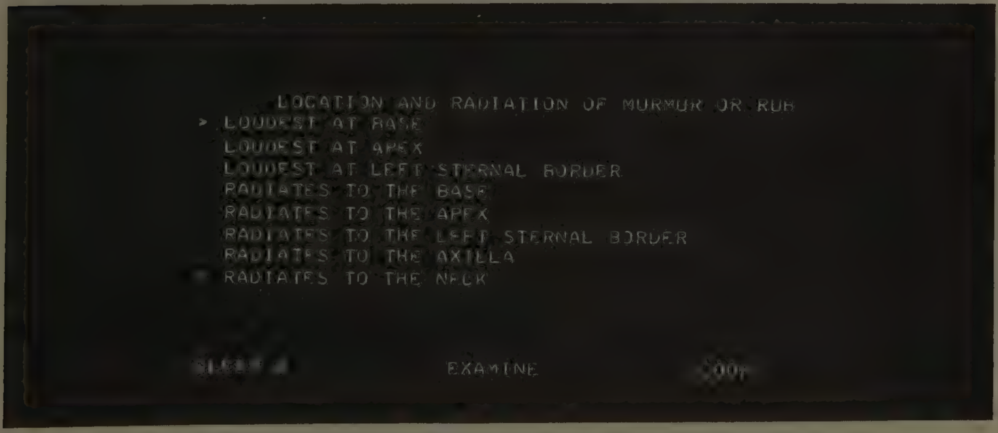
```
                  MURMURS AND RUBS- DESCRIBE [illegible]
            > SYSTOLIC MURMUR
              DIASTOLIC MURMUR
              PERICARDIAL FRICTION RUB




            CLEAR Q              EXAMINE              LOOP
```



```
        CHARACTER OF MURMUR
    EARLY IN CYCLE
    IN MID-CYCLE
    LATE IN CYCLE
    THROUGHOUT CYCLE
  > EJECTION
    WITH THRILL



CLEAR Q              EXAMINE              LOOP
```



```
       GRADE AND QUALITY OF MURMUR
   GRADE I
   GRADE II
   GRADE III
   GRADE IV
   GRADE V
   GRADE VI
   RUMBLING
   BLOWING
   HARSH
   VARYING WITH RESPIRATION
   VARYING WITH POSITION
CLEAR Q                  EXAMINE              LOOP
```



```
       LOCATION AND RADIATION OF MURMUR OR RUB
 > LOUDEST AT BASE
   LOUDEST AT APEX
   LOUDEST AT LEFT STERNAL BORDER
   RADIATES TO THE BASE
   RADIATES TO THE APEX
   RADIATES TO THE LEFT STERNAL BORDER
   RADIATES TO THE AXILLA
   RADIATES TO THE NECK


CLEAR Q               EXAMINE              LOOP
```

Fig. 3 (a, b, c, d, e, f) This series of figures illustrates part of an interactive sequence designed for input of progress notes in a hypertension clinic. Display *a* is presented to the physician, when he has indicated, from a previous display, that he wishes to enter data relevant to chest. The ">" symbol to the left of the fourth choice is displayed as verification by the computer when the physician touches the strip overlying that choice. He has indicated by this action that he intends to discuss the chest examination.

In display *b*, which is presented following display *a*, multiple selections have been entered. Note the three modifiers to the right of the choice list, permitting the physician to indicate that any finding is either normal, unchanged, or abnormal. The symbols ">", "2", or "3" to the left of a particular choice verify that the item has been selected and that a particular modifier pertains (">" is equivalent to "1").

Since the only abnormal finding indicated in display *b* was "heart murmurs or rubs", display c is next presented. The figs. *c-f* show the way in which the physician describes a systolic ejection murmur, grade II, rumbling, loudest at the base, radiating to the neck.

dictating it into a microphone. In the hypertension clinic system, the physician may indicate his desire to dictate by typing a "D" on the keyboard. In this situation, the computer displays a number to be assigned to the particular dictation note. The physician includes this number in the text of his dictation and then continues with the interaction. These dictated comments are later entered into the computer by a secretary typing at a computer terminal. This data is automatically entered into the appropriate part of the patient's file according to the dictation number assigned to it. Dictation is considered to be an important part of the model in evaluating the efficacy of the data structure for recording information and discovering areas in which it is inadequate. The dictation facility is also necessary to provide a flexible output for the physician for recording that information which may be important to his patient's care, but has not been possible to structure. Although any topic can theoretically be structured to whatever degree of depth is desired, there is always a point at which this becomes inefficient and undesirable. Dictation should be a part of this kind of a system, provided that only a small amount of dictation occurs, since it permits the flexibility that makes this system acceptable to the physician. If this requirement is met, one well-trained medical secretary could easily handle the entry of dictated comments by a large number of physicians.

It was shown by the hypertension clinic study that from five to ten experiences in entering progress notes are required before a physician has become sufficiently accustomed to the terminal device and to the organization of the subject material that he is able to use the approach effectively. The majority of this time is spent in learning the organization of the subject material, since the terminal itself is very simple to operate. Two physicians, a clinical fellow, and a senior staff physician, were studied over a four month period, in which they recorded progress notes directly, using the computer system, each on twenty-five unselected patients.

The amount of time required for the physician to enter progress notes was measured as a function of training and familiarity with the system; Fig. 4 illustrates the typical learning curve characterizing each physician's performance. After approximately ten progress notes, both physicians reached a rather stable phase in which only slight additional improvement occurred; for this phase the average time required to record a progress note by computer interaction was found to be less than or equal to that required to write the note by hand. Check lists completed on progress notes written by hand before institution of the computer system were compared with check lists on progress notes generated by computer, with respect to the presence of comments on specific topics considered to be important in progress notes in this clinic. For each of the categories of items checked, the progress notes recorded in the computer were considerably more complete. The physicians seemed to be motivated to write much more explicit notes when using the computer system. In an attitude survey in which they were asked to rate the input method on a scale ranging from "excessively cumbersome" to "very easy to use",

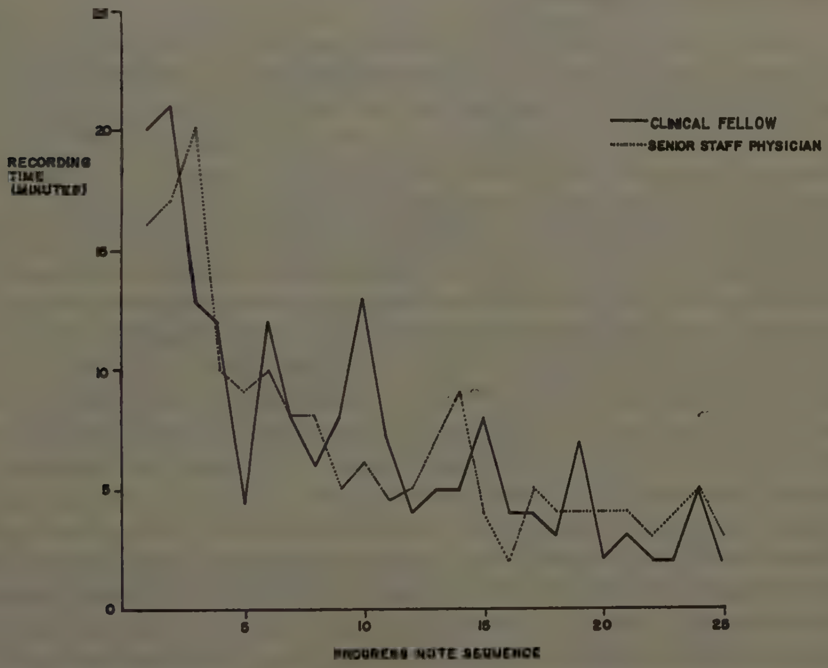


Fig. 4. Time in minutes for clinical fellow and senior staff physician for recording a progress note using the interactive computer system, as a function of training in the use of the system. Note the typical exponential learning curve and the rather similar performance of each physician. After the steep early phase of the curve, for approximately the first ten progress notes recorded by each physician, only slight additional improvement occurred.

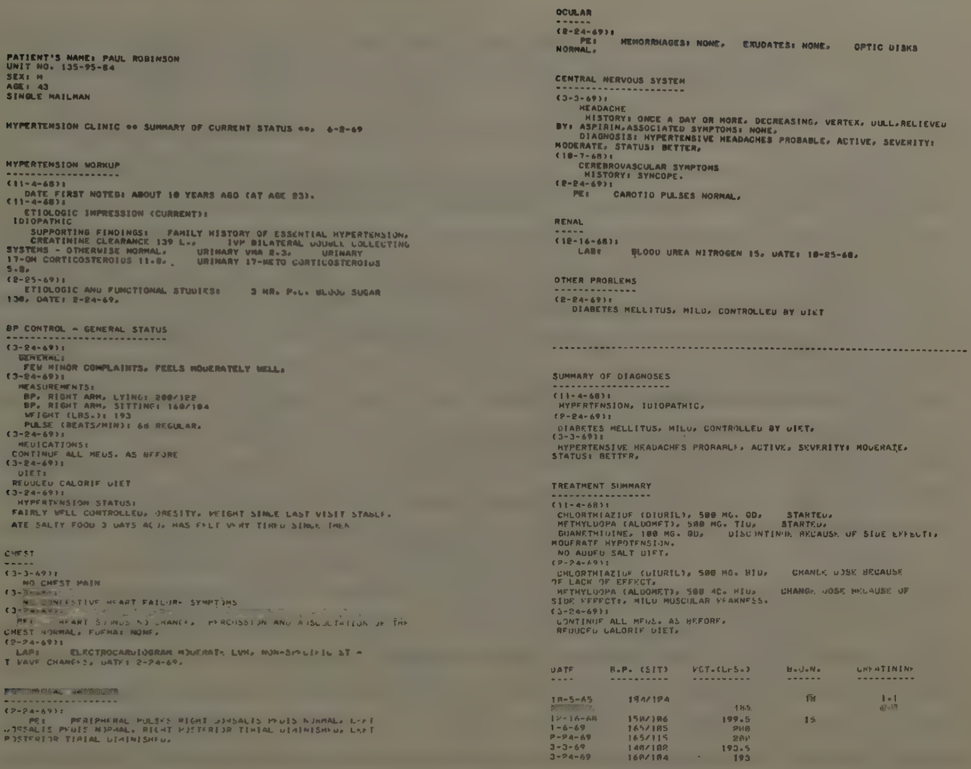
they found it easy to use in almost all cases. They further noted that all or most of the pertinent data could be recorded without dictation, merely by selection from the data structure in all but 12% of progress notes; for these remaining cases, they felt that the dictated comments contributed substantially to the note. Analysis of the dictated comments indicated that an average of 1.2 comments were made per progress note. Ignoring those comments relating to "other problems", which were not intended to be part of the data structure, an average of 0.9 comments were made per patient. It was clear that the privilege of dictation, which was made readily and easily available, was by no means abused.

As these results have indicated, physicians have been willing and able to utilize this kind of approach for the routine recording of clinical information. The training and orientation of physicians in the use of this kind of an approach on a wide scale would optimally occur as a by-product of utilization of such structured input techniques during the process of physician education. In addition, medical schools would need actively to train students in the recording of clinical information.

Summary and Review Capabilities

Following entry of data by a physician, the information exists in a coded form. The coding is hierarchical, with each code reflecting a corresponding level of the data structure associated with the item. A coded version of the progress note may then be stored in the computer file, which contains all previous notes recorded for that patient. This patient file is organized by topic, or problem; within each topic the file is organized chronologically, so that the notes relevant to a particular problem may be traced and retrieved efficiently. When the patient is scheduled for a return visit to the clinic, or when a report is desired for some other reason, a report may be printed by the computer, summarizing the current status of each of the patient's problems. This summary may be formatted by major areas, and within each area by problem, as illustrated by the hypertension clinic summary in Fig. 5. For each problem that has ever been discussed, the date of the last note on that problem is printed, followed by the text of the note itself. The text is regenerated from the coded representation stored in the file. This summary provides an overview of the most recent status of each of the problems presented by the patient. In addition, the computer produces a summary of diagnoses by date, and a summary of all treatment, also by date. The treatment summary includes all changes in medications and reasons for change. Finally, chronologically organized tables of parameters or graphs of pertinent data may be produced. Fig. 6, for example, indicates a bar graph of blood pressure produced with the summary. This summary, printed on three to four pages, may be used by the physician in lieu of the medical record.

During the evaluation of the hypertension clinic system, each physician was given a copy of the summary prior to his interview with and examination of every patient, but was not given the patient's medical record. He was asked to utilize the summary to review his patient's problems, but could review the medical record if he desired. The amount of time required to review the summary was found to be

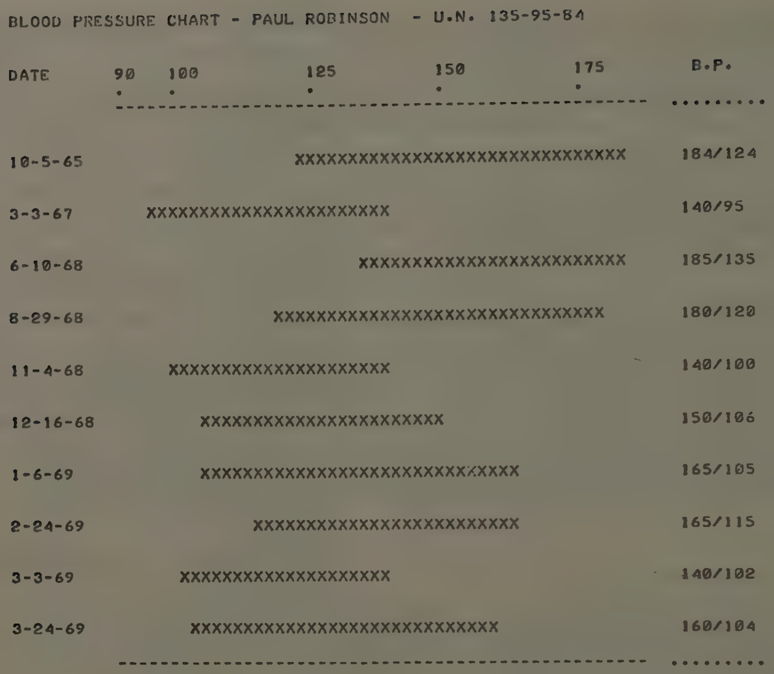


```
PATIENT'S NAME: PAUL ROBINSON
UNIT NO. 135-95-84
SEX: M
AGE: 43
SINGLE MAILMAN

HYPERTENSION CLINIC ** SUMMARY OF CURRENT STATUS **, 6-2-69

HYPERTENSION WORKUP
-------------------
(11-4-68):
   DATE FIRST NOTED: ABOUT 10 YEARS AGO (AT AGE 23).
(11-4-68):
   ETIOLOGIC IMPRESSION (CURRENT):
 IDIOPATHIC
    SUPPORTING FINDINGS:   FAMILY HISTORY OF ESSENTIAL HYPERTENSION,
    CREATININE CLEARANCE 139 L.,     IVP BILATERAL DOUBLE COLLECTING
SYSTEMS - OTHERWISE NORMAL,     URINARY VMA 2.3,     URINARY
17-OH CORTICOSTEROIDS 11.0.     URINARY 17-KETO CORTICOSTEROIDS
5.8,
(2-25-69):
   ETIOLOGIC AND FUNCTIONAL STUDIES:     3 HR. P.C. BLOOD SUGAR
130, DATE: 2-24-69.

BP CONTROL - GENERAL STATUS
---------------------------
(3-24-69):
  GENERAL:
   FEW MINOR COMPLAINTS. FEELS MODERATELY WELL.
(3-24-69):
  MEASUREMENTS:
   BP, RIGHT ARM, LYING: 200/122
   BP, RIGHT ARM, SITTING: 160/104
   WEIGHT (LBS.): 193
   PULSE (BEATS/MIN): 68 REGULAR.
(3-24-69):
  MEDICATIONS:
 CONTINUE ALL MEDS. AS BEFORE
(3-24-69):
  DIET:
 REDUCED CALORIE DIET
(3-24-69):
  HYPERTENSION STATUS:
 FAIRLY WELL CONTROLLED, OBESITY. WEIGHT SINCE LAST VISIT STABLE.
 ATE SALTY FOOD 3 DAYS AGO, HAS FELT VERY TIRED SINCE THEN

CHEST
-----
(3-3-69):
   NO CHEST PAIN
(3-3-69):
   NO CONGESTIVE HEART FAILURE SYMPTOMS
(3-24-69):
  PE:     HEART SOUNDS NO CHANGE,   PERCUSSION AND AUSCULTATION OF THE
CHEST NORMAL, EDEMA: NONE.
(2-24-69):
  LAB:      ELECTROCARDIOGRAM MODERATE LVH, NON-SPECIFIC ST -
T WAVE CHANGES, DATE: 2-24-69.

PERIPHERAL VASCULAR
-------------------
(2-24-69):
    PE:     PERIPHERAL PULSES RIGHT DORSALIS PEDIS NORMAL, LEFT
DORSALIS PEDIS NORMAL. RIGHT POSTERIOR TIBIAL DIMINISHED, LEFT
POSTERIOR TIBIAL DIMINISHED.
```

```
OCULAR
------
(2-24-69):
    PE:    HEMORRHAGES: NONE,   EXUDATES: NONE,    OPTIC DISKS
NORMAL,

CENTRAL NERVOUS SYSTEM
----------------------
(3-3-69):
    HEADACHE
     HISTORY: ONCE A DAY OR MORE, DECREASING, VERTEX, DULL,RELIEVED
BY: ASPIRIN,ASSOCIATED SYMPTOMS: NONE,
    DIAGNOSIS: HYPERTENSIVE HEADACHES PROBABLE, ACTIVE, SEVERITY:
MODERATE, STATUS: BETTER,
(10-7-68):
    CEREBROVASCULAR SYMPTOMS
     HISTORY: SYNCOPE.
(2-24-69):
   PE:    CAROTID PULSES NORMAL.

RENAL
-----
(12-16-68):
    LAB:     BLOOD UREA NITROGEN 15, DATE: 10-25-68,

OTHER PROBLEMS
--------------
(2-24-69):
   DIABETES MELLITUS, MILD, CONTROLLED BY DIET

------------------------------------------------------------------------

SUMMARY OF DIAGNOSES
--------------------
(11-4-68):
 HYPERTENSION, IDIOPATHIC,
(2-24-69):
 DIABETES MELLITUS, MILD, CONTROLLED BY DIET,
(3-3-69):
 HYPERTENSIVE HEADACHES PROBABLE, ACTIVE, SEVERITY: MODERATE,
STATUS: BETTER,

TREATMENT SUMMARY
-----------------
(11-4-68):
 CHLORTHIAZIDE (DIURIL), 500 MG. QD,    STARTED,
 METHYLDOPA (ALDOMET), 500 MG. TID,    STARTED,
 GUANETHIDINE, 100 MG. QD,    DISCONTINUE BECAUSE OF SIDE EFFECT:,
MODERATE HYPOTENSION.
 NO ADDED SALT DIET.
(2-24-69):
 CHLORTHIAZIDE (DIURIL), 500 MG. BID,    CHANGE DOSE BECAUSE
OF LACK OF EFFECT,
 METHYLDOPA (ALDOMET), 500 MG. BID,    CHANGE DOSE BECAUSE OF
SIDE EFFECT:, MILD MUSCULAR WEAKNESS.
(3-24-69):
 CONTINUE ALL MEDS. AS BEFORE,
 REDUCED CALORIE DIET.

DATE       B.P. (SIT)   WGT.(LBS.)     B.U.N.     CREATININE
----       ----------   ----------     ------     ----------

10-5-65     194/124                    [illegible]          1.1
[illegible]                      185                 [illegible]
12-16-68    150/106         199.5        15
1-6-69      165/105           200
2-24-69     165/115           208
3-3-69      140/102         193.5
3-24-69     160/104           193
```

Fig. 5 (a, b) A summary produced prior to the clinic visit on 6-2-69, and reviewed instead of the medical record by the physician. Note that the summary is divided by major sections, as is the progress note, and within each section by topic or problem. The note on each topic is preceded by a date in parentheses, which indicates that date on which the note was entered (the date on which that particular topic was last discussed). Thus the summary represents the most recent status of each topic independently. In addition, the summary contains a chronological list of diagnostic statements made, a summary of treatment, and a table of measurements and laboratory tests by date, of interest to follow-up and management of hypertension.

approximately the same as that required to review a record when no summary was provided. When the summary was available, an overview of all the patient's problems could be had in approximately the same amount of time as that required for the physician to review his last one or two progress notes in the medical record, when no summary was provided. The physicians frequently commented on the usefulness of the summary as a reference while they were talking with the patient. Anecdotally, the summary pointed out a number of circumstances in which laboratory test procedures which were thought to be part of the initial workup of the patient were found to be not actually present in the record, or in which tests of cardiac or renal function had not been ordered in the recent past. They uniformly agreed that the summary could replace the medical record for most problems.



```
BLOOD PRESSURE CHART - PAUL ROBINSON  - U.N. 135-95-84

DATE      90   100          125         150          175         B.P.
          .    .            .           .            .
          --------------------------------------------------  .........

10-5-65                    XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX     184/124

3-3-67       XXXXXXXXXXXXXXXXXXXXXXX                           140/95

6-10-68                          XXXXXXXXXXXXXXXXXXXXXXXXX     185/135

8-29-68                  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX       180/120

11-4-68        XXXXXXXXXXXXXXXXXXXXX                           140/100

12-16-68          XXXXXXXXXXXXXXXXXXXXXXX                      150/106

1-6-69            XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX               165/105

2-24-69                XXXXXXXXXXXXXXXXXXXXXXXXX               165/115

3-3-69          XXXXXXXXXXXXXXXXXXXX                           140/102

3-24-69          XXXXXXXXXXXXXXXXXXXXXXXXXXXXX                 160/104
          --------------------------------------------------  .........
```

Fig. 6 Bar graph of blood pressures by date produced along with summary illustrated in Fig. 5. Range between diastolic and systolic pressure is indicated by X's.

In addition to the summary of current status, the physician is able to use the computer system to selectively review past notes concerning any problem. Using the CRT terminal, he is able to indicate the problem he wishes to review, by pointing to it. The information stored in the patient file for the most recent note on that problem is then retrieved and displayed for him on the scope face, following which he may trace that problem in reverse chronological order, note by note, as far back as information exists. This capability is straightforward once the information is stored in the file organization described above.

Research Retrieval Capabilities

For clinical investigation, the user is able to interact with the computer system in a mode similar to that used for recording information. However, in the retrieval mode, information entered into the system is considered to be an inquiry or retrieval request rather than actual patient data. For example, the physician may enter a statement describing chest pain, frequency more than once a week, lasting more than five minutes, and noted to be relieved by nitroglycerin. In retrieval mode, this is taken as a request to locate all records in the patient file meeting these criteria. Because of the use of a hierarchical coding scheme for data stored in the file, a retrieval request may use as fine a sieve as desired in retrieving clinical records. Because the same interactive system is used to input retrieval requests as to enter data, the user further need never be concerned with the use of codes; all encoding and decoding is an automatic function of the system.

Administrative, Technical, and Economic Considerations

From an administrative or management point of view, an important aspect of an automated medical record system concerns the requirement for security and privacy of clinical information, e.g., determination of access rights to clinical information and methods of enforcement thereof, and the attendant problems generated by production of copies of records (25). Procedures for handling duplicate information, excessive paper, and disposal of this information in a well-controlled manner need to be worked out. In addition, there must be a legally acceptable way for the user to indicate to a computer his authority to either access information or to enter additional information; this identification method must connote authority in the same sense that a signature presently does. The use of identification badges which are machine readable is currently being explored in a number of areas. Another approach consists of the use of passwords or codes, which the user must enter into the computer system to establish his identity.

A major issue in consideration of a computer-based medical record system concerns determination of an appropriate scale for the system. Should systems be located entirely within hospitals, within cities, within regions of the country, or should they be national? This is in part an administrative question concerned with management control of information, but in addition, it is a technological problem which depends primarily upon the availability of economical mass storage of various orders of magnitude. The development of high density, read only optical (laser) memories promises to greatly expand the range of storage capabilities that is economically feasible. Most large hospital record rooms currently use an operational definition of an "active" patient record, such as access within the past few years, to determine whether the record is to be stored in the file room or in a warehouse. In a similar manner, computer-based clinical data management systems will need to support hierarchies of peripheral storage, in which the access time of the storage device utilized is commensurate with the frequency or urgency of the need for retrieval.

It may not be advantageous to set up very large systems, but rather to provide the capability of communicating between one system and another, if the frequency of need for such communication would not be expected to be very often. The other reason for limiting the size of the system concerns the capability of a computer to support multi-terminal access to a central data base, the efficiency with which it does this being inversely proportional to the number of terminals requiring such access. At present, the largest computer systems contemplated might permit 100 to 200 users at terminals to have simultaneous access to the system. More moderate sized computers might support substantially fewer active terminals. It may be assumed that most terminals located in various places in the medical establishment would not be used continuously; provision of a dial-up communication network would permit many more terminals to have access to the system than the system could support simultaneously actually utilizing it.

Because a computer-based medical record system would be used to keep information essential to patient care, reliability of the system is mandatory. The best way to achieve this is to have complete redundancy of all equipment used in the system, so that if any hardware fails, its dual may be put into operation instead. This is another argument for the use of relatively small scale equipment, which might be duplicated in a number of locales. It would then become possible, if any system has failed, to portion the workload among the remaining systems in the area.

The most critical aspects limiting the widespread introduction of an automated medical record system are economic considerations. While the benefits of an ideal medical record system are clear, they cannot be evaluated independently of considerations of cost. Accurate measurements of both benefits and costs are frequently difficult, if not impossible. For example, one of the largest human costs of the present system of recording concerns the frustration that often results from the time spent in locating information. It is likely that this frustration and occasional lack of available information contribute to extra patient days in the hospital, but there is no simple approach to quantitating this. There are at least three different areas in which the relative cost/benefit performance of a manual and an automated medical record system might be evaluated: (1) the time required for the attending physician and hospital staff to record information or to review medical records; (2) the effects on patient morbidity, length of stay in the hospital, and quality of continuing care; and (3) the administrative costs concerned with tracing medical records, pasting and updating records with reports that arrive in the medical record room, pulling and transportation of medical records, weeding of medical record files for inactive records, etc.

While such costs are primarily dependent on personnel time and are therefore continuously increasing, the cost/performance ratio for computers has decreased by a factor of more than ten-fold over the past five or six years. Components of computer systems which are still relatively expensive are random access secondary memory, such as disks, drums, or data cells, and buffered cathode ray tube display terminals; nevertheless both categories of equipment are already beginning to enter a pricing reduction phase much as that which has already occurred for computers.

Conclusion

Pronounced changes in the nature of medical care, its organization, and delivery are beginning to occur. Introduction of computer-based system for clinical information processing can be instrumental in precipitating these changes. Many information handling practices that exist have long been outmoded, and systems such as that described herein have become necessary for medical record management, for patient care, and for research. The model described in this paper has been implemented as a prototype, and evaluated under conditions of operational use. This highly successful but limited experiment has been carried out in order to test some of the basic premises of a system based on physician/computer interaction, and to stimulate further efforts at refinement and extension of the model. As an important aspect of current efforts to study and revitalize our medical care delivery system, careful technical and economic planning with respect to the various considerations raised in this discussion will need to take place. An arduous, coordinated effort of experimentation and development is now necessary, if some of the innovative and hitherto untested approaches to problems of medical record management, which technology can make possible, are to be seriously pursued.

Acknowledgment

The content and philosophy of this article have been significantly influenced by the contributions of the following individuals, to whom the author gratefully extends his appreciation: G. Octo Barnett, M.D., Stuart W. Klein, M.D., George P. Baker, M.D., Anthony G. Oettinger, Ph.D., Jan W. Koch-Weser, M.D., Anthony Robbins, M.D., Jerome H. Grossman, M.D., Roderick E. Prior, James W. Poitras, Edward H. Shortliffe, and Curtis W. Marble.

REFERENCES

1. Drucker, P. F., *The Age of Discontinuity*, Harper and Row, New York, 1968.
2. Mesthene, E. G., How technology will shape the future, *Sci.* **161,** 135-143, 1968.
3. Knowles, J. H., "Medical school, teaching hospital, and social responsibility", in *The Teaching Hospital* (Knowles, J. H., ed.), Harvard Univ. Press, Cambridge, Massachusetts, 1966.
4. Jydstrup, R. A. and Gross, M. J., Cost of information handling in hospitals, *Health Serv. Res.* **1,** 235-271, 1966.
5. Slack, W. V., Hicks, G. P., and Reed, C. E., A computer-based medical history system, *New Eng. J. Med.* **274,** 194-198, 1966.
6. Mayne, J. G., Weksel, W., and Sholtz, P. N., Toward automating the medical history, *Mayo Clin. Proc.* **43,** 1-25, 1968.
7. Grossman, J. H., Barnett, G. O., McGuire, M. T., Greenes, R. A. and Swedlow, D. B., Evaluation of computer acquisition of patient medical histories. *Clin. Res.,* 1969 (abstract).
8. Lindberg, D. A., Collection, evaluation, and transmission of hospital laboratory data, *Meth. Inform. Med.* **6,** 97-107, 1967.
9. Pribor, H. C., Kirkham, W. R., and Hoyt, R. S., Small computer does a big job in this hospital laboratory, *Mod. Hosp.* **110,** 104-107, 1968.
10. Hicks, G. P., Gieschen, M. M., and Slack, W. V., Routine use of a small digital computer in the clinical laboratory, *JAMA* **196,** 973-978, 1966.
11. Katona, P. G., Pappalardo, A. N., Marble, C. W., Barnett, G. O. and Pashby, M. M., Automated chemistry laboratory: application of a novel time-shared computer system, *Proc. IEEE* **57,** 2000-2006, 1969.
12. Korein, J., Bender, A. L., Rothenberg, D., and Tick, L. J., Computer processing of medical data by variable-field-length format, *JAMA* **186,** 132-138, 1963.
13. Levy, R. P., Cammarn, M. R., and Smith, M. J., Computer handling of ambulatory clinic records, *JAMA* **190,** 1033-1037, 1964.
14. Yoder, R. D., Preparing medical record data for computer processing, *Hospitals* **40,** 75-76, 1966.
15. Collen, M. F., Periodic health examinations using an automated multitest laboratory, *JAMA* **195,** 830-833, 1966.
16. Hall, P., Mellner, C., and Danielsson, T., J5-A data processing system for medical information, *Meth. Inform. Med* **6,** 1-6, 1967.
17. Barnett, G. O., and Greenes, R. A., Interface aspects of a hospital information system, (presented at Conf. on Data Mechanization and Computers in Clinical Medicine, Jan., 1968), *Ann. N.Y. Acad. Sci.* **161,** 756-768, 1969.
18. Kiely, J. M., Juergens, J. L., Hisey, B. L., and Williams, P. E., A Computer-based medical record, *JAMA* **205,** 571-576, 1968.
19. Slack, W. V., Peckham, B. M., and Van Cura, L. J., A computer-based physical examination system, *JAMA* **200,** 224-228, 1967.
20. Weed, L. L., Medical records that guide and teach. *New Eng. J. Med.* **278,** 593-599, 652-657, 1968.
21. Greenes, R. A., Barnett, G. O., Klein, S. W., Robbins, A., and Prior, R. E. Recording, retrieval, and review of medical data by physician/computer interaction, *New Eng. J. Med.* **282.**
22. Greenes, R. A., *A Computer-Based System for Medical Record-Keeping by Physicians* Ph.D. Thesis, Harvard University, 1969.
23. Weed, L. L., personal communication, Jan., 1969.
24. Gordon, B. L., Biomedical language and format for manual and computer applications, *Dis. Chest* **53,** 38-42, 1968.
25. Freed, R. N., A legal structure for a national medical data center, *AFIPS Conf. Proc.,* Vol. 33 (FJCC), Thompson Book Co., Washington, D. C., 1968, pp. 387-394.

COUNTWAY LIBRARY
HC 2461 %

QT 34 I647 1969a
Advances in medical physics; sympos
Countway Library AMA4687
3 2044 016 421 307

Date Due

1 ~~NOV '72 CLW~~

~~DEC 1 2~~ 1973

~~MAY 2 5 1976~~

~~JAN 1 0 1983~~

Demco 38-297

QT 34 I647 1969a
Advances in medical physics; sympos
Countway Library AMA4687
3 2044 016 421 307